MONOGRAPHS

JOURNAL OF THE NATIONAL CANCER INSTITUTE

NATIONAL CANCER INSTITUTE

1996
Number 20

Quality of Life in Clinical Cancer Trials

Contents

(Contents continued on back cover)

MONOGRAPHS
JOURNAL OF THE NATIONAL CANCER INSTITUTE

Number 20 *1996*
ISSN 1052-6773

EDITORIAL POLICY: Manuscripts from key conferences dealing with cancer and closely related research fields, or a related group of papers on specific subjects of importance to cancer research, are considered for publication, with the understanding that they have not been published previously and are submitted exclusively to *Journal of the National Cancer Institute Monographs*. All material submitted for consideration will be subject to review, when appropriate, by at least one outside reviewer and one member of the Editorial Board of the *Journal of the National Cancer Institute*. Opinions expressed by the authors are not necessarily those of the publisher or the editors.

Proposals for monographs should be submitted to the Editor-in-Chief, *Journal of the National Cancer Institute*, National Cancer Institute, 9030 Old Georgetown Rd., Bethesda, MD 20814.

Journal of the National Cancer Institute Monographs are available on request to members of the National Cancer Institute Information Associates Program (one copy of each monograph per member). Monographs are also available through the U.S. Government Printing Office. To request a copy of a monograph, or for more information about the Information Associates Program, call 1-800-624-7890 (301-496-7600 outside the United States).

Quality of Life in Clinical Cancer Trials

Proceedings of a Workshop
Held at the
National Institutes of Health
Bethesda, Maryland
March 1-2, 1995

Sponsors

National Cancer Institute, Division of Cancer Treatment, Diagnosis, and Centers and Division of Cancer Prevention and Control

Workshop Faculty

Frank Baker
Julie Beitz
Andrew S. Bradlyn
Pennifer Erickson
Clare Gnecco
Carolyn C. Gotay
Jimmie C. Holland
Penelope Hopwood
Gwendoline M. Kiebert
Edward L. Korn
David Machin
Mary S. McCabe
Carol M. Moinpour
David Osoba
Leslie Robison
Edward Trimble
Claudette G. Varricchio
Richard B. Warnecke
Jane Weeks

Contents

Introduction

*Claudette G. Varricchio, Mary S. McCabe, Edward Trimble, Edward L. Korn**

In July 1990 (*1*), the National Cancer Institute (NCI) held a quality-of-life (QOL) meeting (*a*) to define elements of QOL that are relevant to clinical decision making and serve as end points in cancer clinical trials, (*b*) to evaluate currently available instruments for QOL assessments and strategies for implementation, (*c*) to identify site-specific questions of high priority, and (*d*) to examine issues regarding the integration of findings from therapeutic evaluations and QOL measurements.

The meeting reported in this monograph, "Workshop on Quality of Life in Clinical Cancer Trials," represents the next step in advancing QOL research in NCI trials through an assessment of the progress that has been made in NCI-sponsored QOL research during the last 5 years. The general goal of this meeting was to focus on and to refine expectations for QOL research so that essential evaluations can be done in QOL as an addition to other clinical information, particularly in a time of limited resources. The specific focuses of the current workshop were (*a*) to re-evaluate clinical research areas in which QOL questions are a priority, (*b*) to address issues of implementation and collection of QOL data in NCI-sponsored clinical trials, and (*c*) to focus on new methods in QOL research, such as outcome studies and methods of assessing QOL in culturally diverse populations.

The meeting participants included the QOL researchers from the NCI's Cooperative Groups and the Community Clinical Oncology Program (CCOP) research bases, representatives from the British Cancer Research Campaign, the Medical Research Council (U.K.), NCI-Canada, and the European Organization for Research and Treatment of Cancer. As a formal part of the meeting, participants were asked to review current QOL research in their respective groups and to present the groups' plans for future QOL investigations.

This monograph presents the papers from the meeting with selected discussion. A summary of NCI-sponsored QOL protocols from the groups is included along with the QOL instruments used to address the research questions. When available, citations of published findings are listed.

Many issues were explored in the papers presented. Some of these issues were as follows: Is QOL needed in every trial? What are the advantages and disadvantages of including QOL? Given the resources needed for QOL studies, how can groups best set priorities on the use of group resources? Since inclusion of women and minorities is mandated in the National Institutes of Health (NIH) Revitalization Act of 1993 (Public Law 103-43) (*2*), how do the groups plan to comply with the NIH guidelines? What is the best way to develop and refine measurement scales, especially disease-specific modules and cultural adaptations/translations that are valid and reliable? These issues require further thought and investigation.

Other areas of interest or speculation came from the discussions. They included the manner in which QOL research is presented and the forums where results are presented. Inclusion of QOL results reported with clinical end points often leads clinicians to question the rationale for QOL research and the clinical application of the findings. The question "What are QOL data and how does one use them?" must be answered eloquently and cogently. The question of multiple measures in trials versus standard measures used across trials was raised. Is comparability across trials warranted at this point in the development of QOL measures? Is there a need for depth (disease-specific question) in measurement as well as a need for breadth (all domains of QOL) in conceptual issues? A goal must be to understand not only who are doing better, but also why they are doing better. Criteria for excluding a subject from a trial must be looked at closely and must be justified if the trials are to be representative and generalizable. How does one design QOL studies to be inclusive? Special populations include those with linguistic and cultural diversity, persons with low literacy ability, children, the elderly, and hearing-impaired or visually impaired persons. What should the sample size be for QOL findings to be meaningful? Is it the same, larger, or smaller than that needed to answer a treatment question? Operational issues are of concern to the groups who are faced with the pragmatic reality of limited resources and manpower. These questions will no doubt need to be addressed through the conduct of trials.

The results of QOL assessment in clinical trials should focus on interventions to lessen the negative impact of cancer and its treatment on QOL; these interventions must build on the descriptive QOL research findings. QOL will continue to expand in clinical cancer research, such as prevention trials in which risk assessment (e.g., genetic or environmental exposure) and notification methods are developed and tested. There will be a need for assessment of the effect of knowledge of risk status on the person's QOL. The translation of QOL findings into valid, effective clinical applications is the most important concern of researchers and clinicians at this time. It is important

---

**Affiliations of authors:* C. G. Varricchio (Division of Cancer Prevention and Control), M. McCabe, E. Trimble, E. L. Korn (Division of Cancer Treatment Diagnosis, and Centers), National Cancer Institute, Bethesda, MD.

Correspondence to: Claudette G. Varricchio, D.S.N., R.N., National Institutes of Health, 6130 Executive Blvd., MSC 7340, Bethesda, MD 20892-7340.

See "Note" section following "References."

that QOL research continues to develop and address cancer research questions of clinical importance to patients.

References

(*1*) Nayfield SG, Hailey BJ, McCabe M. Quality of life assessment in cancer clinical trials: report of the Workshop on Quality of Life Research in Cancer Clinical Trials, July 16-17, 1990. Bethesda (MD): US DHHS, 1991.

(*2*) Federal Register Vol. 59, No. 59, March 28, 1994.

Note

Workshop supported by the Division of Cancer Prevention and Control and Division of Cancer Treatment, Diagnosis and Centers, National Cancer Institute.

Trial-Related Quality of Life: Using Quality-of-Life Assessment to Distinguish Among Cancer Therapies

*Carolyn Cook Gotay**

Issues in selecting quality-of-life (QOL) measures that are best suited to assessing differences among treatments in cancer clinical trials, as well as challenges to interpreting QOL outcome data, are discussed. When used in the context of randomized trials of cancer therapies, QOL assessments must provide an answer to the question, "Did the treatments differentially affect patient well-being?" In order to detect differences in treatment efficacy against a background of great similarity, the broad concept of QOL needs to be refined to reflect "trial-related QOL." In many cases, this will entail emphasis on actual patient experience of symptoms and functional changes, as opposed to emphasis solely on evaluation and satisfaction. A model is proposed to identify cognitive, emotional, and sociocultural factors that influence a patient's QOL evaluation and that need to be considered in understanding the meaning of QOL data. [Monogr Natl Cancer Inst 1996;20:1-6]

The potential contributions of quality-of-life (QOL) data to cancer therapy evaluation are increasingly recognized. Only a handful of phase III clinical trials that include results of the QOL assessments have been published to date (*1,2*). Active portfolios of trials including QOL outcome measures, however, are maintained by all of the multisite clinical cooperative groups in the United States (*3*), as well as trial groups in Canada (*4*), Europe (including the U.K.) (*5*), and Australia (*6*). The results of a number of these ongoing trials will begin to be available in the next few years. For example, the results of the first Southwest Oncology Group QOL studies will be reported at the plenary session of the October 1995 group meeting (Moinpour C: personal communication, 1995).

With increased interest in QOL assessment and concomitant resources devoted to this activity come increased expectations for the contributions of QOL data to interpreting trial data. Some of these expectations may not be met. In the enthusiasm for using QOL measures, limitations to their interpretation have not been fully considered. This article discusses issues in selecting QOL measures that are best suited to assessing differences between treatments in cancer clinical trials, as well as challenges to interpreting QOL outcome data. Discussion will focus on the purpose of assessing QOL in a given trial, specific challenges to QOL assessment posed by randomized trials, and difficulties in interpreting QOL data. We will propose a model to explain QOL and to identify needs for additional research.

For What Purpose Is QOL Being Assessed in a Given Trial?

There are numerous reasons why QOL assessment might be included in a particular cancer treatment study (*7-9*). Cella and Tulsky (*9*) have provided a parsimonious taxonomy of purposes for measuring QOL: 1) to identify the full range of side effects and impacts of the treatments in order to assess rehabilitation needs, 2) to compare treatments in a trial, and 3) to use QOL ratings as a predictor of response to future treatment. Data collected for the first purpose can identify patients at risk to provide supportive interventions and to modify treatment regimens. Data collected for the second purpose can be used to determine which treatment should be the standard of care. With regard to the third purpose, base-line QOL scores can serve either as a prognostic indicator or as a basis for stratification in the random assignment of patients to treatments.

These purposes cannot necessarily be achieved through the same approach to measurement. For example, documenting the impact of treatments may require a comprehensive assessment that includes questions about patient experience in the multiple dimensions that make up QOL. Virtually all researchers agree that QOL involves a number of relatively independent domains, including, at a minimum, physical, functional, psychological, and social well-being. Some researchers also emphasize other areas, such as symptoms, sexuality, spiritual concerns, and satisfaction with health care (*10*). A broad and comprehensive approach to assessment is likely to be particularly useful for treatments for which little is known about potential effects on patient well-being. New treatments may have an impact in areas that are not expected by the investigators. In one of the earliest studies in this area, Sugarbaker et al. (*11*) demonstrated that radiation therapy used in limb-sparing procedures had an unanticipated negative impact on patient sexual functioning. For distinguishing among treatments, the same broad approach to assessing QOL that is useful in understanding patient experience may not provide the specific information that is necessary to distinguish between treatments. This point will be discussed in more detail in the next section.

**Correspondence to:* Carolyn Cook Gotay, Ph.D., University of Hawaii Cancer Research Center, 1236 Lauhala St., Honolulu, HI 96813.
See "Notes" section following "References."

The use of QOL data for prognosis or stratification is sufficiently recent that the best approach to assessment has not yet been identified. Several studies (*12,13*) have shown that patient-rated overall QOL assessments predict survival better than physician-rated performance status. Coates et al. (*14*) made a head-to-head comparison of patient and physician QOL ratings. Breast cancer patients participating in a clinical trial of chemotherapy completed a QOL questionnaire (including questions on physical well-being, mood, pain, nausea and vomiting, and appetite as well as an overall rating), and their physicians also completed the Quality of Life Index (QLI), a multidimensional QOL assessment (*15*), and a performance status measure. Results showed that patient ratings of physical well-being (but not overall QOL) and physician-rated QLI were both statistically significant and independent predictors of length of survival. This study points out the need for additional research to identify the best approach to QOL assessment for prognostic purposes (*16*).

It should not be inferred from the above discussion that QOL assessment can be used for one purpose and one purpose only in a given study. Often there are multiple reasons why QOL assessment should be conducted and several different ways this information can be applied. QOL assessment, however, generally takes place in the context of limited resources. Not only are the data management and analysis capabilities of the cooperative groups limited, but also the patient's ability to provide information may be limited by fatigue and motivation. Priorities need to be set to ensure that the appropriate QOL data are available to address the study purpose. If resources are sufficient to permit additional data collection, such data are invariably likely to provide useful information for patient care. At a minimum, however, the researcher needs to be certain to collect data appropriate to study the hypotheses.

What Are Constraints to Measuring QOL in Clinical Trials When the Goal Is to Distinguish Among Treatments?

It is reasonable to consider that one of the primary reasons that QOL assessment has been accepted and adopted in therapeutic and drug development research derives from the current status of clinical research in cancer. For many cancers, there have not been major improvements in therapeutic cure rates since the success of chemotherapeutic agents in the 1960s. Many patients are living longer, however, even if they ultimately die of their disease. In this context, additional measures of treatment efficacy, such as QOL assessments, assume increased importance in determining standards of care in cancer treatment and the approval of new pharmaceutical agents.

Phase III trials are the gold standard for evaluating new cancer treatments. In phase III trials, patients are randomly assigned to one of two or more treatments, generally including the standard "best treatment" and a new treatment that is believed to be at least as good and perhaps better. Eligibility criteria are used to ensure patient safety, as well as to restrict participation to a well-defined group of patients for whom treatment effects may most likely be detected. As a result of these eligibility restrictions, most participants in phase III clinical trials constitute a selected and nonrepresentative group of patients; as a consequence of randomization, the patients in all treatment arms should be equivalent on any important variables related to outcomes except for the treatment they receive.

The implications of this design are that detecting differences in QOL between treatment arms is apt to be very difficult. The patients reflect great similarity in their disease status on entry to the study, since site and stage of diagnosis will be identical across treatments. In addition, many aspects of the treatments will be identical. For example, monitoring schedules, tests performed, symptom control regimens, and numerous other aspects of treatment are specified and controlled in the study protocol. In addition, the majority of phase III studies currently ongoing in the cooperative groups involve comparisons of different chemotherapeutic regimens. This is also true for many phase II studies, especially those that test new drugs; QOL assessment may also be considered in these studies, as witnessed by the active interest and participation of the U.S. Food and Drug Administration and pharmaceutical industries in this field (*17*).

This situation clearly poses challenges for QOL assessment, since a QOL measure would need to be very sensitive to detect differences in treatment efficacy against a background of great similarity in patient populations. The ability to detect differences between identical patient groups and/or among treatment regimens that are alike on many dimensions requires focused QOL assessment strategies.

Most investigators in the cancer field restrict their definition of QOL in clinical trials to health-related QOL (HRQOL). Specifically, most investigators agree that it makes sense to limit the QOL end point of interest in a study of a treatment intervention or in a population with compromised health status such as cancer patients to the dimension(s) that are likely to be affected by the intervention or health status of the patient. As a result, there may be aspects of QOL that are very important in an individual's "subjective evaluation of life as a whole" [a frequently cited definition of QOL offered by DeHaes (*18*)] that are excluded from consideration when HRQOL is assessed. These aspects of QOL include dimensions such as the environmental quality, physical safety of the neighborhood, and quality of public schools. While these are critical aspects (and in fact are key components of comparative ratings of QOL in other contexts, such as comparing QOL in cities across the country), they are outside the realm of being affected by treatment interventions or health status.

We would like to propose that a similar "funneling" occurs in selecting measures of QOL used in clinical trials of cancer treatment in order to develop an assessment of trial-related QOL (TRQOL). Specifically, when one wishes to detect differences between two or more treatment arms, a number of the domains commonly included in QOL assessments are unlikely to be differentially affected by the treatments under study. For example, a comparison of two chemotherapies may be unlikely to have different impacts on spiritual concerns and family functioning. At the same time, the treatments might differ in, for example, their effect on sleep patterns or fatigue. Since these two areas are not assessed beyond a single question (if that) on most HRQOL questionnaires, standard tools would be unlikely to be sensitive enough to show differences among treatments if in-

deed they existed. Given that there is a limit as to how many questions can be included in a given trial, researchers need to ensure that they cover the critical aspects of TRQOL. If they are fortunate enough to have the resources to assess HRQOL, or even QOL, they will gain additional valuable information. (In fact, virtually no research has investigated the relationship between overall QOL and HRQOL. These data would help to put into perspective the relative weight patients attribute to their health concerns in the context of their life as a whole.) However, when QOL assessment is included in order to distinguish among treatments, the most important objective is to be able to answer the question, "Did the treatments differentially affect patient well-being?" The specific assessments made need to be sufficiently sensitive to answer this question.

What Are Difficulties in Interpreting QOL Findings?

Most of the models of QOL that have been presented to date consist of identifying different dimensions that may influence QOL. However, little attention has been directed toward specifying the relationships between symptoms and functioning, performance and satisfaction, occurrence of a symptom and experience of the symptom as a problem, and most of the other concepts that are loosely described as relating to QOL. Similarly, virtually no attention has been given to identifying variables that predict QOL, apart from determining if QOL varies as a function of treatment. However, the question "What factors are associated with high levels of QOL?" remains unanswered.

Part of the difficulty in exploring such a question derives from the way that many researchers define QOL. Most discussions of cancer-related QOL stress its subjective nature, emphasize that QOL can be assessed only from the perspective of the patient, and stress that the patient's evaluation is the "gold standard" (*19*). However valid this approach may be for understanding an individual patient, it is difficult to accept as a criterion for success of cancer treatment. Consider the following case:

> Mr. G. is an 87-year-old prostate cancer patient. He lives with his wife of 58 years and his daughter and her family in a comfortable home. In addition to his cancer diagnosis, he has several other comorbid diseases, which have left him with one leg amputated above the knee, a brain tumor, a weakened right side, paralysis of half his face, and an eye sewn shut. In addition, he is half-blind in his other eye. He is largely confined to a hospital bed. When he completed the QOL questionnaire and was asked, "Do you have any trouble taking a short walk outside the house?" he responded. "No. If someone helps me out of bed, and I use my two prostheses, and a couple of canes, I can walk." When he was asked to rate the overall quality of his life from 1 to 10, he selected 10 and remarked, "I have a wife that's the tops, two daughters who are quite successful, and my mother and dad were really great. My quality of life is hard to beat, because I've been getting everything I wanted, as far as I'm concerned, and no problems." [From (*20*), cited with patient permission]

It is clear that the patient's evaluation of his QOL as "excellent" and "a 10" is a candid and accurate reflection of his perspective. Efforts to modify treatments to mitigate symptoms or enhance other outcomes, however, are still to be strived for if possible, despite the patient's satisfaction and experienced high levels of QOL. Relying completely on patient evaluations without equal attention to more objective aspects of well-being limits the usefulness of QOL data as an outcome measure in cancer treatment. It is clear that there are easier ways to make patients happy than by giving them cancer therapy.

In addition, Mr. G's excellent QOL was certainly influenced by his personal and social resources, as well as his own values and attitudes. Multiple perceptual, motivational, and external factors intervene between an experience related to cancer and/or its therapy (such as a symptom or a change in functioning) and its evaluation by the patient as a problem or an effect on QOL. All of these variables are largely outside the realm of cancer therapy. Models are needed to identify and link independent variables to patient-assessed QOL. Such models will facilitate the development of interventions that incorporate individual patient factors with QOL assessments.

Fig. 1 presents a model that attempts to identify factors that may affect a patient's evaluation of HRQOL as related to cancer. This schema elaborates on models presented by Selby (*2*) and Wilson and Cleary (*21*). The model assumes that the patient who is asked to provide a rating of his or her HRQOL engages in a multistep process. Psychological (including cognitive and emotional factors) and sociocultural filters affect how the patient experiences and assesses the effects of cancer diagnosis and treatment.

The model begins when cancer is diagnosed and treated. As a consequence of the disease and/or treatment, the patient may experience symptoms and functional limitations. This is one juncture when data assessing patient experience provide direct information about whether or not various symptoms are experienced as well as the functional limitations that may result.

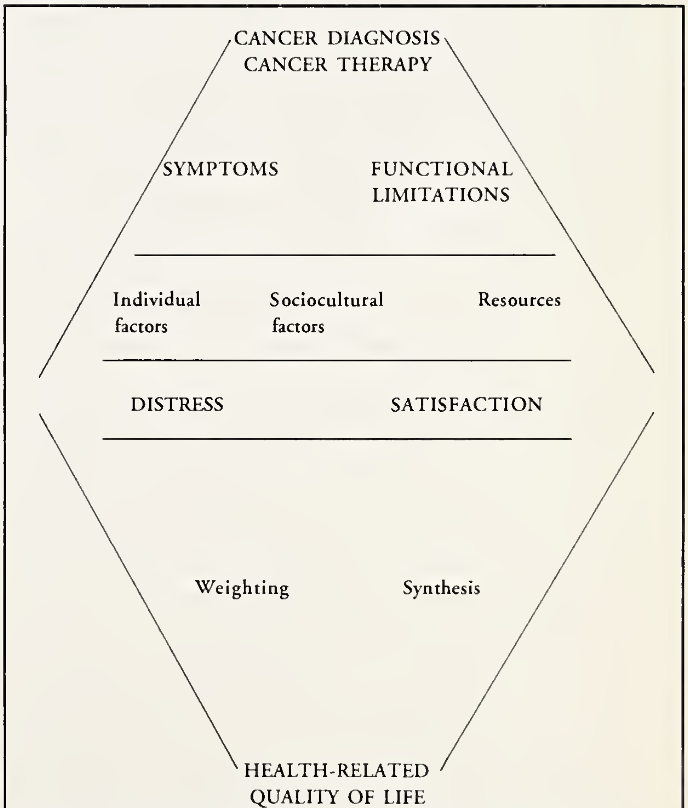

Fig. 1. Conceptual model of health-related quality of life.

Measures of functional capacity—whether it is possible for a patient to perform a defined task—as opposed to assessment of performance during everyday activities are more likely to yield data that reflect the effects of treatment as opposed to motivations and lifestyle. (Measures of actual functioning rather than capacity may provide more useful information to guide individual patient rehabilitation.)

A number of available QOL assessment questionnaires do not emphasize patient experience; instead, they ask directly about patient evaluations. The specific questions vary according to QOL questionnaire. For example, the QOL Index of Ferrans and Powers (*22*) asks cancer patients to indicate their satisfaction and rated importance of aspects of functioning, while the majority of questions in the CARES (*23*) address the degree to which cancer patients have difficulty in different areas. However, whether or not symptoms or functional changes result in distress (or possibly have a positive impact) depends on other variables. These variables include individual factors, such as the patient's motivation, interpretation, expectations, and personality.

With respect to motivation, some people attempt to supersede potential limitations and find ways to surmount the challenges they face (such as Mr. G. described above); such people would probably not experience distress to the same degree as others. The role that motivation can play in whether a disability becomes a handicap is well recognized in the rehabilitation literature (*24*).

An individual's interpretation of symptoms and other limitations also affects the degree of distress that is experienced. For example, patients may be willing to experience many objectively negative side effects of treatment (e.g., chemotherapy-associated emesis) on a short-term basis if they believe that they are going to get better as a result of the treatment. This kind of patient interpretation may help to explain the finding of Coates et al. (*25*) that QOL, as well as tumor response, was better for patients who were given continuous rather than intermittent chemotherapy. Treatment was administered to the patients in the continuous-therapy arm until their disease progressed; hence, receiving treatment may have signified to them that they were doing well, which in turn may have engendered more favorable QOL ratings. This explanation is conjectural, since Coates et al. did not collect data about patient interpretation, but it offers one way to explain somewhat puzzling findings.

Expectations may have a great deal to do with whether or not an individual can accept limitations. For patients who expect to be "back to normal," a functional limitation may be much more distressing than for individuals who expect that they might need to live with limitations. The relationship between experience and expectations has been identified as a key determinant of QOL by Calman (*26*) and Cella and Cherin (*27*).

Many other individual differences may affect the degree to which distress is experienced. For example, there is evidence to support the existence of a dispositional complaining style. Some people express dissatisfaction across situations and would be expected to report lower levels of well-being regardless of the circumstances (*28*). Premorbid health conditions also need to be considered. For example, the importance of considering previous psychopathology in understanding dysfunctional coping with cancer has been amply demonstrated (*29*). Comorbid physical problems also affect health ratings. For example, in our experience assessing QOL, patients frequently make comments such as, "I have pain, but I don't know if it's because of the cancer or my arthritis." Collecting data about concurrent and previous health problems would aid in untangling the impact of cancer and cancer therapy from more general concerns.

Sociocultural factors also have an influence on whether or not a particular symptom or functional disability gives rise to patient-rated distress. Perhaps the clearest illustration can be found with respect to pain, where cultural variation has been extensively studied. While the ability to detect pain stimuli appears to be equivalent across cultures (*30*), both the meaning and expression of pain are culturally bound (*31*), as research during the past 40 years has demonstrated [e.g., (*32-34*)]. One of the earliest studies in this area (*34*) showed that while "old Americans" (U.S.-born Anglos of third or greater generation status) tended to be stoic and unemotional in their responses to pain, Italian-American and Jewish patients were expressive and verbal in communicating discomfort. Such ethnocultural differences could have a major effect on QOL ratings.

The resources possessed by the patient constitute another category of individual factors that may influence whether a symptom leads to distress. Economic resources are the most straightforward. For example, if a patient can pay for someone to obtain groceries and clean the house, functional limitations may not be as distressing. Social support is another important resource. A third kind of resource is whether appropriate health care has been provided. For example, a patient may be suffering from depression in response to the cancer diagnosis and be receiving psychoactive medications or other support by his or her physician. The resultant QOL rating may reflect no mood dysfunction because the symptom has been adequately remediated. However, the need to inquire about such matters in the course of assessing HRQOL has not been discussed in the literature. In addition, existing QOL questionnaires do not build in questions about whether a patient is currently receiving support to mitigate symptoms or functional limitations.

There is yet a further evaluative step that patients need to take when they make an overall assessment of their HRQOL. Although it is an unconscious process for most patients, they need to make a number of judgments in order to render an overall QOL rating. Weights must be assigned according to the subjective importance of different dimensions of well-being, experienced distress must be multiplied by these weights, and these factors need to be synthesized in order to determine a final answer. Depending on individual values, the same distress ratings may give rise to different overall HRQOL scores.

The strength of assessing global HRQOL is its emphasis on the importance of individual differences in QOL and the need to consider each patient's perspective (*35*). At the same time, this poses a considerable difficulty when one is attempting to draw conclusions about HRQOL as a function of cancer treatment. If QOL is completely subjective and is affected by a multiplicity of factors well outside the jurisdiction of influence by cancer therapy, then it seems a very difficult task to detect differences due to treatment. Random assignment of patient to treatment condition should ensure that the conditions are balanced; i.e., in-

dividual variations in patient motivations, values, and the like should be equally represented across treatments. Such variables, however, may give rise to so much error in outcome measurement that differences cannot be detected.

The fact that studies have found differences between treatment groups in randomized studies indicates that, despite the considerable individual variation among patients, some differences among treatments are apparently large enough that variation in HRQOL can be detected. Since individual patient values, personality, and so forth are less likely to have been affected by the treatment, such differences in treatments, however, are likely to stem from variation in symptoms and/or functioning. [As we proposed earlier in discussing the study by Coates et al. (*25*), it is possible that individual factors such as expectations may be differentially affected by treatment arm.] By the same token, some treatments with considerable differences in symptomatology and function have not been demonstrated to have a consistent and demonstrable difference in HRQOL. Consider the impact of mastectomy versus conservative surgery for breast cancer, where the confluence of evidence points to no consistent HRQOL advantage for either treatment (*36*). In this instance, it is likely that individual patient variables (such as the importance of physical appearance), as well as the impact of a diagnosis of cancer apart from treatment, mediate differential HRQOL ratings. Attention to factors like those represented in Fig. 1 will aid in interpreting the findings, either differences between treatments or lack thereof, and avoid coming to an erroneous conclusion such as "the kind of surgery received for breast cancer does not affect QOL."

What Are Implications for Future Research?

Research assessing QOL as a consequence of cancer has made enormous strides in the past decade. HRQOL has been recognized and incorporated as an end point in trials of therapy, procedures have been developed to ensure quality control of the data in multisite studies, and questionnaires to assess HRQOL have been developed and continue to be validated. Several areas, however, deserve additional attention to ensure that HRQOL data fulfill their potential.

1) *Basic research is needed to understand more fully the contributions of HRQOL data to cancer therapy evaluation.* The relationship between HRQOL data and other measures needs to be clarified to identify the distinct contributions of HRQOL data. For example, what are the relationships among toxicity ratings, symptoms, functioning, and HRQOL?

Consider toxicity ratings and HRQOL. We would not expect perfect correlations between patient HRQOL ratings and clinician-rated toxic effects, based on demonstrated differences between patient and observer ratings (*37*). However, do toxicity ratings demonstrate differences among treatments in the same direction and of the same magnitude as HRQOL ratings? How much additional predictive validity do HRQOL data provide over and above toxicity ratings? Is it possible for toxicity and HRQOL to be noncorrelated or even negatively correlated? The cooperative groups maintain careful records of an extensive battery of toxic effects, which could be compared with patient ratings in those studies that include QOL assessments. This kind of analysis could be relatively easily accomplished and would constitute an important contribution to the field by indicating areas where detailed patient reports are most critical.

2) *HRQOL measurement needs to be tailored to the study purpose.* Given resource constraints, an assessment strategy that addresses multiple purposes may not be possible. For studies in which HRQOL is used to distinguish among cancer treatments, the assessment tool must be sensitive and focused enough to detect small differences against a background of considerable similarity in patients and, frequently, treatment regimens. This requirement may necessitate the use of TRQOL (trial-related QOL) questionnaires in addition to (preferably) or instead of more standard assessment tools.

3) *Attention needs to be directed at assessing more objective effects of cancer diagnosis and treatment, in addition to patient evaluation.* The majority of questionnaires that have been developed to measure HRQOL in cancer are heavily weighted to patient evaluations of distress or satisfaction. However, as we have discussed, patient ratings of distress are affected by many variables that fall well outside the scope of cancer therapy. It is recommended that investigators adopt HRQOL measures that include assessment of symptoms and functional deficits as well as patient-evaluated distress and problems. In fact, the most commonly used measure in health assessment outside of cancer is the Sickness Impact Profile (*38*), a scale that emphasizes functional assessment. HRQOL assessments in cancer patients should ensure the inclusion of questions that address functional status (*39,40*).

In addition, consideration needs to be given to alternative approaches to HRQOL assessment. Most assessment to date has utilized patient self-reports. Self-reports, however, are no less subject to methodological biases than other sources of data. Each approach to data collection has its own distinct strengths and limitations, and a triangulation approach, which relies on the convergence of data from different sources (*41*), is the optimal approach when possible. Alternative sources of data, such as observer reports, medical records, and behavioral ratings, should be considered in addition to self-reports (*42*); see (*43*) for a useful example of a behaviorally based scale assessing several aspects (speech and eating behavior) of HRQOL particularly important for head and neck cancer patients.

4) *Models of QOL need to be developed and tested.* In order to understand why a patient experiences high or low levels of QOL, influences on this evaluation need to be identified and quantified. Little attention has been directed at identifying factors that influence QOL ratings and which cancer patients experience notably high or low HRQOL. As Till (*44*) pointed out, ultimately, "it may be much more important to try to understand and learn from the process used by the patient to construct his or her report than to obtain the outcome of the process." The model outlined in this article was presented to stimulate thinking and empirical efforts toward this goal.

References

(*1*) Osoba D. Lessons learned from measuring health-related quality of life in oncology [*see* comment citation in Medline]. J Clin Oncol 1994;12:608-16.

(*2*) Selby P. Measurement of quality of life in cancer patients. J Pharm Pharmacol 1993;45 Suppl 1:384-6.

(3) Nayfield SG, Ganz PA, Moinpour CM, Cella DF, Hailey BJ. Report from a National Cancer Institute (USA) workshop on quality of life assessment in cancer clinical trials. Qual Life Res 1992;1:203-10.
(4) Osoba D. The Quality of Life Committee of the Clinical Trials Group of the National Cancer Institute of Canada: organization and functions. Qual Life Res 1992;1:211-8.
(5) Maguire P, Selby P. Assessing the quality of life in cancer patients. Br J Cancer 1989;60:437-40.
(6) Olweny CL. Quality of life in cancer care. Med J Aust 1993;158:429-32.
(7) Gotay CC, Korn EL, McCabe MS, Moore TD, Cheson BD. Quality-of-life assessment in cancer treatment protocols: research issues in protocol development. J Natl Cancer Inst 1992;84:575-9.
(8) Moinpour CM, Feigl P, Metch B, Hayden KA, Meyskens FL Jr, Crowley J. Quality of life end points in cancer clinical trials: review and recommendations. J Natl Cancer Inst 1989;81:485-95.
(9) Cella DF, Tulsky DS. Quality of life in cancer: definition, purpose, and method of measurement. Cancer Invest 1993;11:327-36.
(10) Donovan K, Sanson-Fisher RW, Redman S. Measuring quality of life in cancer patients. J Clin Oncol 1989;7:959-68.
(11) Sugarbaker PH, Barofsky I, Rosenberg SA, Gianola FJ. Quality of life assessment of patients in extremity sarcoma clinical trials. Surgery 1982;91:17-23.
(12) Ruckdeschel JC, Piantodosi S, the Lung Cancer Study Group. Quality of life assessment in lung surgery for bronchogenic carcinoma. J Thor Surg 1991;6:201-5.
(13) Ganz PA, Lee JJ, Siau J. Quality of life assessment. An independent prognostic variable for survival in lung cancer. Cancer 1991;67:3131-5.
(14) Coates A, Gebski V, Signorini D, Murray P, McNeil D, Byrne M, et al. Prognostic value of quality-of-life scores during chemotherapy for advanced breast cancer. Australian New Zealand Breast Cancer Trials Group [*see* comment citation in Medline]. J Clin Oncol 1992;10:1833-8.
(15) Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, et al. Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. J Chronic Dis 1981;34:585-97.
(16) Weeks J. Quality-of-life assessment: performance status upstaged? [editorial; comment] [*see* comment citations in Medline]. J Clin Oncol 1992;10:1827-9.
(17) Johnson JR, Temple R. Food and Drug Administration requirements for approval of new anticancer drugs. Cancer Treat Rep 1985;69:1155-9.
(18) de Haes JC. Quality of life: conceptual and theoretical considerations. In: Watson M, Greer S, Thomas C, editors. Psychosocial oncology. Oxford: Oxford Univ Press, 1988:61-70.
(19) Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements [*see* comment citations in Medline]. JAMA 1994;272:619-26.
(20) Gotay CC. Quality of life in cancer patients in Hawaii. Ongoing study funded by the National Cancer Institute, 1995.
(21) Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995; 273:59-65.
(22) Ferrans CE, Powers MJ. Quality of life index: development and psychometric properties. ANS Adv Nurs Sci 1985;8:15-24.
(23) Ganz PA, Schag CA, Lee JJ, Sim MS. The CARES: a generic measure of health-related quality of life for patients with cancer. Qual Life Res 1992;1:19-29.
(24) Wright BA. Physical disability: a psychological approach. New York: Harper & Row, 1960.
(25) Coates A, Gebski V, Bishop JF, Jeal PN, Woods RL, Snyder R, et al. Improving the quality of life during chemotherapy for advanced breast cancer. A comparison of intermittent and continuous treatment strategies. N Engl J Med 1987;317:1490-5.
(26) Calman KC. Quality of life in cancer patients—an hypothesis. J Med Ethics 1984;10:124-7.
(27) Cella DF, Cherin EA. Quality of life during and after cancer treatment. Compr Ther 1988;14:69-75.
(28) Watson D, Clark LA. Negative affectivity: the disposition to experience aversive emotional states. Psychol Bull 1984; 96:465-90.
(29) Andersen BL. Psychological interventions for cancer patients to enhance quality of life. J Consult Clin Psychol 1992;60:552-68.
(30) Trill MD, Holland JC. Cross-cultural differences in the care of patients with cancer. Gen Hosp Psychiatry 1993;15:21-30.
(31) Kleinman A. Culture, the quality of life and cancer pain: anthropological and cross-cultural perspectives. In: Ventafridda V, van Dam F, Yancik R, Tamurini M, editors. Assessment of quality of life and cancer treatment. Washington, DC: Elsevier Science Publishers, 1986:43-50.
(32) Bates M, Edwards WT. Ethnic variation in the chronic pain experience. Ethn Dis 1992;2:63-83.
(33) Gaston-Johansson F, Albert M, Fagan E, Zimmerman L. Similarities in pain descriptions of four different ethnic-culture groups. J Pain Symptom Manage 1990;5:94-100.
(34) Zborowski M. Cultural components in response to pain. J Soc Issues 1952;8:16-30.
(35) Ganz PA. Quality of life and the patient with cancer. Individual and policy implications. Cancer 1994;74(4 Suppl):1445-52.
(36) Kiebert GM, de Haes JC, van de Velde CJ. The impact of breast-conserving treatment and mastectomy on the quality of life of early-stage breast cancer patients: a review. J Clin Oncol 1991;9:1059-70.
(37) Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992;45:743-60.
(38) Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.
(39) Guyatt GH, Cook DJ. Health status, quality of life, and the individual [comment] [*see* comment citations in Medline]. JAMA 1994;272:630-1.
(40) Spitzer WO. State of science 1986: quality of life and functional status as target variables for research. J Chronic Dis 1987;40:465-71.
(41) Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959;56:81-105.
(42) Gotay CC. Patient-reported assessments versus performance-based tests. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. New York: Raven Press. In press.
(43) List MA, Ritter-Sterr C, Lansky SB. A performance status scale for head and neck cancer patients. Cancer 1990;66:564-9.
(44) Till JE. Quality of life measurements in cancer treatment. In: DeVita VT Jr, Hellman S, Rosenberg SA, editors. Important advances in oncology 1992. Philadelphia: Lippincott, 1992:189-204.

Notes

Supported in part by Public Health Service grants CA61711 and CA01642 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

I gratefully acknowledge the assistance of Mary Clarke in interviewing cancer patients.

Quality-of-Life End Points in Cancer Clinical Trials: The U.S. Food and Drug Administration Perspective

*Julie Beitz, Clare Gnecco, Robert Justice**

Increasingly, quality-of-life (QOL) end points are being incorporated into randomized, controlled clinical trials in oncology. The Oncologic Drugs Advisory Committee (U.S. Food and Drug Administration) has recommended that beneficial effects on QOL and/or survival be the basis for approval of new anticancer drugs. Therefore, from a regulatory standpoint, for drugs that do not have an impact on survival, demonstration of a favorable effect on QOL is more important than most other traditional measures used to assess efficacy, such as objective tumor response. Trials incorporating QOL questions will be evaluated on the basis of how well they address the stated objectives. The clinical protocol should delineate investigators' hypotheses and choice of validated instruments and should specify a detailed statistical analysis plan describing strategies for handling missing data. The U.S. Food and Drug Administration welcomes the opportunity to explore with investigators the use of QOL instruments in the design of cancer clinical trials. [Monogr Natl Cancer Inst 1996;20:7-9]

The Oncologic Drugs Advisory Committee (*1*) of the U.S. Food and Drug Administration has recommended that beneficial effects on quality of life (QOL) and/or survival be the basis for approval of new anticancer drugs. Therefore, from a regulatory standpoint, for drugs that do not have an impact on survival, demonstration of a favorable effect on QOL is more important than most other traditional measures used to assess efficacy, such as objective tumor response.

Methods to Assess QOL

A spectrum of QOL instruments has been developed, ranging from global to disease-specific to ad hoc instruments that are specific to a single study (*2*). In the past, the ad hoc approach has dominated QOL assessment in clinical trials evaluating new anticancer agents. These instruments often lack rigorous validation and do not allow for cross-study comparisons.

Global instruments, designed for use across a wide range of chronic disease populations, are most applicable to health policy research. Their advantage is in examining a wide range of potential impacts of disease on mental functioning and social functioning. When applied in oncology settings, however, these instruments may fail to address important issues relevant to the cancer patient (i.e., side effects of anticancer treatment or tumor-related symptoms) and may lack sensitivity to changes in important but localized aspects of QOL.

Disease-specific instruments, on the other hand, overcome the problems inherent in the ad hoc approach, have the advantage of addressing problems specific to a given cancer patient population, and may permit cross-study comparisons. These instruments allow separation of favorable and unfavorable events—they don't just give a "score." Thus, in the context of a cancer clinical trial, the impact of a new therapy on the individual QOL components of interest can be weighed separately. This is potentially most useful when the new anticancer therapy offers no real improvement in survival or cure rate as compared with standard therapies.

To balance the trade-offs inherent in the global and disease-specific instruments, many experts have proposed synthesizing these approaches, with the use of a cancer-specific core instrument (a more focused type of global instrument), supplemented with disease-specific or treatment-specific assessment modules. Examples include the European Organization for Research and Treatment of Cancer core QOL questionnaire (QLQ-C30), supplemented by a lung cancer-specific module (QLQ-LC13), or the general Functional Assessment of Cancer Therapy Scale (FACT-G), supplemented by the ovarian cancer-specific module (FACT-Ovarian) (*3,4*).

What Trial Designs Are Appropriate for Prospective QOL Assessment?

Certainly, the phase III randomized clinical trial is the obvious venue for QOL assessment, given that the findings of such trials will likely have an impact on future clinical practice, and these trials generally enroll large numbers of patients who are followed for extended periods. Most importantly, these trials allow valid use of a highly subjective instrument.

A host of logistical issues present themselves that can have a major impact on the integrity of QOL assessments. It is critical that assessments be performed at times when patients can be

**Affiliations of authors:* J. Beitz, R. Justice (Division of Oncology and Pulmonary Drug Products, Office of Drug Evaluation I), Clare Gnecco (Division of Biometrics, Office of Epidemiology and Biostatistics), U.S. Food and Drug Administration, Rockville, MD.

Correspondence to: Julie Beitz, M.D., U.S. Food and Drug Administration, 5600 Fishers Lane, HFD-150, Rockville, MD 20857.

See "Note" section following "References."

most objective about their QOL. Careful consideration should be given to the length of the instrument, especially for the seriously ill patient or in longitudinal studies that stipulate frequent assessments. If feasible (e.g., in the comparison of oral agents), double blinding is preferable. If not, study personnel directly involved in the QOL assessment should be blinded to patients' treatment assignments and to their responses to treatment, even though this situation may be difficult. Finally, feedback from the investigator or his/her staff that systematically influences patients' sense of well-being should be avoided, as this is one of the major sources of bias in open-label trials (*5*).

Other clinical trial designs that have incorporated QOL instruments include phase II and even selected phase I trials. Proponents favor this approach, since early patient perspectives on a new anticancer therapy may have an impact on its future development. Moreover, experience with QOL instruments early on may allow further refinement prior to large-scale use in later phase trials. Any open (uncontrolled) QOL assessment presents problems, however; at present, the impact of early incorporation of QOL assessments on the approval of new anticancer agents or on acceleration of approval remains to be determined (*6*).

QOL instruments may further our knowledge of the impact of chemoprevention. Large multicenter trials to address this issue are currently under way. They involve subjects at increased risk for the development of breast or prostate cancer. The usefulness of QOL data in the approval of new chemopreventive agents is not yet known, but whenever large numbers of disease-free people are exposed to a therapy, it is worth looking for subtle adverse effects.

QOL assessments may further our knowledge of the impact of cancer and its treatment on selected populations, such as the elderly, cancer survivors, and family members of cancer patients. Future studies evaluating the impact of genetic risk notification will likely include QOL assessments. While information obtained from such studies may not have a direct impact on the approval of a new anticancer therapy or of an existing one for a new indication, it may prove clinically useful to practicing physicians and their patients.

Lessons Learned From the Review of QOL Studies

First and foremost, it is critical that investigators identify the purpose of the effort and the nature of the problem they wish to address. Although QOL is potentially an issue in any clinical trial, it is not feasible to measure QOL in all trials. Thus, investigators must decide whether it is truly necessary to obtain QOL data in a given clinical setting. If it is deemed desirable to study QOL, then investigators must decide which domains (functional, psychologic, social, or somatic) are of greatest interest. Again, it may not be feasible to measure all domains.

Next, investigators must select from the host of QOL instruments those that have been validated for the population of interest and that measure the desired end points. If valid instruments do not yet exist for a given patient population (e.g., pancreatic cancer patients), it may be necessary to pilot one or more instruments prior to initiating a large, complex, controlled trial.

Investigators must determine who should conduct each QOL instrument, when each should be administered, and how the objectivity and consistency of patient responses will be ensured. Careful consideration must be given to the statistical analysis of the often-voluminous amount of QOL data collected.

Statistical Design and Analysis Issues

Several important elements should be addressed at the design stage in protocols for studies with QOL end points. They include the following: 1) Prospective specification should be made of a small number of the most important hypotheses of interest; 2) supporting documentation must be obtained regarding the validation of the psychometric properties of the instrument to be used for the disease indication under study; and 3) since oncology trials are often unblinded, it is important to demonstrate symptomatic improvement or some other objective evidence of a beneficial treatment effect (e.g., increased objective response of adequate duration) in addition to instrument-assessed QOL improvement. The disease-specific portion of the instrument is most directly related to the strength of evidence the U.S. Food and Drug Administration will utilize in evaluating claims of significant improvement in QOL. The global portion, of course, serves an important role not only in providing an overall profile but also in serving as a consistency check. More specific QOL protocol design guidelines include the following: 1) provide a detailed schema delineating exactly at what time periods the instrument will be administered, in addition to providing justification for how appropriate these intervals are; 2) state the personnel who will be responsible for administering the instrument and the type of training they will receive or similar information if the instrument is self-administered by the patient; 3) address how bias will be minimized; 4) state what steps will be taken to avoid missing data and how this situation will be handled, should it occur; 5) avoid highly correlated questions; and 6) provide detailed, concomitant medication logs.

Careful prospective planning at the design stage can substantially reduce problems in analyzing QOL study data. The types of problems the U.S. Food and Drug Administration has encountered in the past include the following: 1) sizable amounts of missing data, particularly base-line data, which can render an analysis meaningless; 2) failure to adjust for the confounding effects of concomitant medications, e.g., antidepressants; 3) improper imputation for missing values; and 4) inappropriate analytic methods, e.g., multiple assessments over time and multiple end points (subsets of the scale) without correcting for multiple analyses, and failure to adjust for a patient's base-line status.

Analysis strategies for dealing with missing data include LOCF (last observation carried forward) and end-point analysis in which only two values per patient are used, viz., the base-line value and the last value recorded. Both of these strategies are highly dependent on the assumption of missing mechanisms (i.e., data missing completely at random, missing at random, or missing because of informative censoring). Serious bias can occur if missing data are related to a nonrandom mechanism. Averaging or prorating may not always be appropriate, since such strategies assume that all questions in a domain have equal

weight; this may not always be tenable. In addition, if only certain responses are missing, the concern arises as to whether these were just inadvertent omissions or whether the patient found the question intrusive or objectionable in some way. Thus, an integral part of any QOL analysis should include a thorough investigation of the missingness pattern by treatment arm.

For statistical analysis of QOL data, the U.S. Food and Drug Administration suggests some general guidelines to drug sponsors. In addition to univariate analyses and graphic displays, a strategy should be provided that investigates the temporal element. Acceptable methods include repeated measures ANCOVA (analysis of covariance), if the proper assumptions are met, or a formal longitudinal analysis such as a linear mixed effects model and/or GEE (general estimating equation) approach if ANCOVA is not justified. It is very important to provide a detailed QOL analysis plan in the protocol with the following elements: 1) missing value strategy, 2) details of how the missingness pattern will be investigated and dealt with in the analysis, and 3) full details of the statistical methodology to be used. When such specifications are prospectively provided, the U.S. Food and Drug Administration can comment on and assist with analysis strategies.

Conclusions

Interest in QOL assessments in clinical cancer research has been growing rapidly. Inclusion of QOL end points, particularly in phase III randomized trials, will likely be the rule, rather than the exception, for the foreseeable future. Thus, although QOL analyses to date have been inadequate to support drug approval, data from clinical trials incorporating QOL questions will likely be utilized by sponsors to strengthen applications of new anticancer agents or new indications of existing anticancer agents to the U.S. Food and Drug Administration.

Data from trials incorporating QOL instruments will be evaluated on the basis of how well they address the objectives as stated prospectively in the clinical protocol. Specific QOL questions should be chosen carefully. Use of unvalidated instruments or inappropriate analytic methods will likely be challenged. Missing values, either at base line or as a result of patients dropping out, and improper missing value imputation will seriously hamper the interpretation of QOL data. Given the unique challenges that face investigators in the development of clinical trials in oncology, the U.S. Food and Drug Administration welcomes the opportunity to discuss with them the use of QOL measures in the drug development process.

References

(*1*) Johnson JR, Temple R. Food and Drug Administration requirements for approval of new anticancer drugs. Cancer Treat Rep 1985;69:1155-9.

(*2*) Aaronson NK. Methodologic issues in assessing the quality of life of cancer patients. Cancer 1991;67(3 Suppl):844-50.

(*3*) Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. Eur J Cancer 1994;30A:635-42.

(*4*) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.

(*5*) Nayfield SG, Hailey BJ. Quality of life assessment in cancer clinical trials: report of the Workshop on Quality of Life Research in Cancer Clinical Trials. Bethesda (MD): National Cancer Institute, July 1990.

(*6*) Shoemaker D, Burke G, Dorr A, Temple R, Friedman MA. A regulatory perspective. In: Spilker B, editor. Quality of life assessments in clinical trials. New York: Raven Press, 1990:193-201.

Note

We thank Dr. Robert Temple for his critical reading of this manuscript.

Costs of Quality-of-Life Research in Southwest Oncology Group Trials

*Carol M. Moinpour**

Quality-of-life (QOL) research in Southwest Oncology Group (SWOG) trials has achieved increasing support over the past 5 years. The purpose of this paper is to estimate the cost of performing QOL research in SWOG trials. During the month of January 1995, we tracked staff time expended for QOL tasks at the SWOG's Operations Office and Statistical Center. Of interest was a description of average costs per patient enrolled in existing SWOG trials (both open and closed), including protocol development, ongoing data monitoring, and QOL data analysis. The findings emphasize the personnel-intensive nature of this research and highlight the role of "start-up" costs, especially in terms of programmer time. It is estimated that average monthly direct costs associated with implementing a QOL study and monitoring and analyzing QOL data over the life cycles of current and closed SWOG QOL protocols are $7304; a $443 per QOL patient total cost figure is also presented. Costs associated with initiating QOL research in cooperative groups are substantial (4-5-year start-up investment) but are expected to decline after systems for monitoring, retrieving, and analyzing QOL data are in place. Funding issues are addressed. [Monogr Natl Cancer Inst 1996;20:11-6]

There has been increasing interest in documenting the effect of cancer treatment on patient functioning (*1*). Quality-of-life (QOL) end points have been added to the battery of traditional clinical end points, such as tumor response and survival (*2-9*). Weighting survival time with QOL has been proposed using methods such as QALYs or Quality-Adjusted Life Years (*10*) and QTWiST or Quality-adjusted Time Without Symptoms of disease and Toxicity of treatment (*11*). In addition, QOL data have been used to predict patient survival (*12-14*), to suggest interventions for cancer survivors (*15,16*), and to provide ongoing monitoring of patient functioning outside the clinical trials setting (*17*). The National Cancer Institute (NCI) has supported QOL research through its Division of Cancer Prevention and Control (DCPC), Division of Cancer Treatment (DCT), and the Surveillance, Epidemiology, and End Results Program (*5,6,18*); methodologic support for QOL research has also been provided (*19*). The Food and Drug Administration has allowed submission of QOL data as part of the review process for new oncologic drugs (*20*) but has shown more interest in symptom status and physical functioning dimensions, particularly the extent to which such data document improvement in secondary end points, such as tumor response.

From the start of QOL research in cancer clinical trials, investigators recognized that inclusion of QOL assessments required attention to quality control. Aaronson et al. (*21,22*) described methodologic difficulties encountered in the collection of QOL data for European Organization for Research and Treatment of Cancer (EORTC) trials, noting that the clinical feasibility of QOL studies (e.g., the number of assessments) must be carefully considered prior to implementation; they also addressed the problem of missing data when patients become too ill to complete questionnaires (*22*). The National Cancer Institute of Canada (NCIC) has been particularly successful in implementing quality control procedures and maintaining excellent questionnaire submission rates over time in its clinical trials (*1,23*). A primary reason for the NCIC's success is the centralized monitoring system that tracks and reminds institution staff about the scheduled QOL assessments. Another method for successfully minimizing missing data is the centralized telephone assessment approach used in Cancer and Leukemia Group B (CALGB) trials (*24*). In its first QOL study, the Southwest Oncology Group (SWOG) experienced such poor questionnaire submission rates that the QOL assessments in a breast cancer trial were terminated (*25*). As a result of this experience, the SWOG developed a set of policies to guide the assessment of QOL in selected trials (*2*). The SWOG's assessment guidelines have been previously described (*2,26,27*).

The purpose of this paper is to describe central office costs of doing QOL research in SWOG trials. At the NCI meeting summarized in this issue, the author was asked to describe how to do QOL research in cooperative groups without breaking either the "back" or the "bank" of the cooperative group mechanism. This request was thoughtfully considered in the context of what appears to be declining funds for QOL research in cooperative groups. We were aware that very little was known about how much it did cost to add QOL end points to cancer clinical trials. Those conducting QOL research in cooperative groups have increasingly recognized the substantial time required to develop QOL protocols, to develop and implement quality control systems, and to tailor analysis techniques for QOL data. The following cost estimates describe staff hours required for central office processing of QOL studies and place a dollar figure on this level of effort. Cost estimates will be presented in terms of

**Correspondence to:* Carol M. Moinpour, Ph.D., Southwest Oncology Group Statistical Center, MP-557, Fred Hutchinson Cancer Research Center, 1124 Columbia St., Seattle, WA 98104.

See "Notes" section following "References."

average cost per patient in SWOG protocols involving QOL end points; cost estimates cover the level of effort across the total time period of these trials.

Methods

Cost estimates in this paper address primarily personnel time and salaries of SWOG Statistical Center and Operations Office staff who design QOL studies, monitor data collection, and analyze QOL data. The cost figures represent a cross-sectional estimate of costs for ongoing QOL research in SWOG trials. That is, for 1 month, we examined Statistical Center and Operations Office activities involved with all ongoing and closed (data analysis under way) studies with QOL end points. Although any single month could reflect idiosyncratic fluctuations in staff time, we believe that the three main types of activity, 1) protocol and questionnaire development, 2) implementation and conduct of the monitoring system, and 3) data analysis, were reasonably captured in January 1995. Since we just now have completed trials with QOL end points, use of earlier periods would not have allowed an estimate of average analysis time and would have been less representative of the full range of required QOL activities.

Table 1 shows studies with QOL end points that required staff effort in January 1995, providing information relevant to our method of estimating costs; for example, the number of QOL patients registered in 1994 is provided and, for comparison, those registered over the period 1990 through 1994. The activation and closing dates for the trials are important because the trials vary in how long they are open and require Statistical Center attention. However, the activation and closing dates in Table 1 do not reflect the substantial amount of time prior to trial activation associated with protocol development and programming and data analysis time following trial closure. This time is captured in the estimates of staff QOL time during January 1995 for such trials as SWOG-9327 (not open) and SWOG-9346 (May 15, 1995, opening). Two closed trials received attention for data cleaning and analysis during January 1995.

To obtain estimates presented in Table 2, key Statistical Center staff kept logs that tracked how much time each spent on QOL data during the month of January 1995. Personnel time in Table 2 is based on a working month of 173.3 hours (average working hours per month for 1995).[1] Tables 3 and 4 attach cost estimates to the time and effort described in Table 2. Direct costs for the QOL full-time equivalent (FTE) employee effort in Table 1 are presented in Table 3. Salaries include a range of fringe benefits (22.5% to 28%) based on the type of position. In most cases, an average salary for a staff type is used (e.g., average of Statistical Center Data Coordinator salaries). Costs are in 1994 dollars. An attempt is also made to estimate the cost of basic operating resources expended for the conduct of QOL research at the Statistical Center. For 1994, we determined that of the 5174 patients registered in all SWOG trials, 331 (6%) of these patients were registered to QOL studies; 6% was then applied to each of seven categories of operating expenses (Table 3). Estimates of SWOG QOL costs presented at the NCI March meeting excluded SWOG patients registered in trials coordinated by other cooperative groups because the work associated with such patients was much less than that required for patients in SWOG-coordinated trials. However, quality control procedures for SWOG patients registered to non-SWOG trials have recently been upgraded so that data monitoring for SWOG- and non-SWOG-coordinated trials is now more similar. Therefore, the current calculations include SWOG patients registered to non-SWOG trials containing QOL assessments.

Table 4 extends QOL cost estimates to total (i.e., direct plus indirect) costs. Costs are in 1994 dollars; the rate for indirect costs is 70%. The primary summary variable is assessment-related costs per patient registered in QOL studies (i.e., monthly QOL direct or total costs per monthly QOL registrations). We are attempting to show the additional cost per patient of adding QOL assessments to clinical trials. Based on Table 1, an average of 28 QOL registrations was used for monthly estimates (331 patients registered to QOL studies in 1994/12). The estimate of cost per QOL patient is not for a single trial but captures the workload associated with current and closed SWOG trials involving QOL end points. The estimate is not an annual cost of QOL research because it reflects costs associated with ongoing studies over the life of these trials.

Results

Table 2 indicates the personnel-intensive nature of this work. The resource demands follow from maintaining the same quality control standards for QOL data as those maintained for the clinical database. However, although a large proportion of Statistical Center staff time is devoted to the design and implementation of quality control procedures, protocol development and data analysis require substiantial staff resources. This can be seen in

Table 1. Studies with QOL end points, 1990-1994

Study No.	Disease site	Phase	Date activated (closed)	No. of patients 1994	No. of patients 1990 through 1994
SWOG-8994*	Stage C prostate cancer	III	2/15/90 (———)	27	145
SWOG-9039*,†	Stage D2 prostate cancer	III	10/1/90 (9/15/94)	108	714
SWOG-9045*,†	Advanced-stage colorectal cancer	III	3/1/91 (12/31/93)	—	287
SWOG-9021‡	Brain metastases	II	7/15/91 (12/1/94)	14	47
SWOG-9248	Metastatic breast cancer	II	5/15/93 (2/1/94)	20	125
SWOG-9235	Advanced-stage prostate cancer	II	12/15/93 (6/1/94)	48	53
SWOG-9208*	Early stage Hodgkin's disease	III	4/15/94 (———)	8	8
SWOG-9324§	Relapsed ovarian cancer	II	3/15/95 (———)	—	—
SWOG-9346§	Stage D2 prostate cancer	III	5/15/95 (———)	—	—
SWOG-9327§	Advanced-stage (cachexia)	II (Randomized)	———	—	—
Total SWOG QOL patients				225	1371
SWOG patients in other group QOL studies				106	128
Total QOL patients‖				331	1499
Total patients¶				5174	33 402

*Companion study to therapeutic protocol.

†Data analysis tasks in January 1995.

‡Therapeutic trial closed early because of accrual problems.

§Protocol and forms development/quality control system tasks in January 1995.

‖Includes SWOG patients registered to SWOG QOL trials, patients registered to SWOG-coordinated QOL trials by other cooperative groups, and SWOG patients registered to other cooperative group QOL studies.

¶Includes phase I, II, III, and other (e.g., cancer control) trials. Patient registrations for SWOG-coordinated trials provided by other cooperative groups are included for the reason outlined in footnote ‖. Prostate Cancer Prevention Trial registrations have been excluded.

Table 2. QOL personnel: tasks and hours/month

Required staff (% FTE)*		Tasks
Operations Office Personnel (San Antonio, TX), protocol coordinator	(15%)	Prepares/distributes protocols, amends protocols, fields inquiries
Statistieal Center Personnel (Seattle, WA)		
Ph.D. Psychologist	(25%)	Coordinate QOL research, work with Behavioral and Health Outcomes Committee, conduct QOL training, establish centralized monitoring procedures for QOL questionnaires, develop and review QOL questionnaires and forms, review protocol design, estimate sample size, develop analysis plan, retrieve and manipulate data, conduct analyses, produce QOL sections of semiannual report of studies, prepare and review QOL manuscripts
M.S. Biostatistician	(10%)	
Ph.D. Biostatisticians	(13%)	
Total	(48%)	
Programmers	(17%)	Prepare forms, write data entry programs, generate QOL assessment calendars for each newly registered patient to be sent to institutions, prepare programs for entry and quality control checks of QOL data, develop data dictionaries for QOL data, generate tables in clinical database for QOL data, write programs to monitor overdue QOL questionnaires (SWOG Expectation Report), write program to extract calendars and QOL expectations reports for QOL study coordinators, write programs to retrieve QOL data for analysis, assist with QOL section of semiannual report of studies
Data technicians	(24%)	Open, sort, file, and mail QOL forms to QOL study coordinators, enter QOL data, correct database for amended QOL forms
Data coordinator	(12%)	Reviews protocols and new forms, develops eligibility checklists for QOL studies, sends expectation reports to institutions, reviews charts to resolve missing QOL data, contacts institutions regarding missing QOL data, fields inquiries Re: QOL studies, assists with data cleaning for QOL analyses

*Personnel time based on working month of 173.3 hours.

Table 2, where QOL work consumes almost one-half FTE per month of primarily Ph.D.-level staff.

Another nontrivial resource use in Table 2 deals with data processing (24% FTE per month) and programming (17% FTE per month). Programmer time for QOL data represents a substantial upfront cost, probably for 4-5 years. QOL data must be integrated into an existing clinical database system but the fit, although initially not good, can be achieved over time. The SWOG's experience has shown that programmers need ample lead time prior to activation of a new protocol to handle the QOL programming tasks. Most QOL programming tasks are more complicated in varying degrees than those associated with traditional clinical data because of the nature of the QOL data. For example, a data dictionary requires more text to describe both variables and response options, and programming for QOL database retrieval is different from that used for clinical data. A new QOL programming task is incorporation of QOL data in the SWOG's Expectation Report, a computer-generated report sent to an investigator indicating that data are overdue for a particular patient. Each follow-up QOL assessment must be programmed as a separate expectation variable to be appropriately resolved when overdue forms have been submitted. Initially, and to some degree continually, the Expectation Report is demanding of programmer time. However, the Expectation Report has become an important component of the SWOG's centralized monitoring system for QOL data, since resolution of missing data must occur in order for an institution to remain in good standing in SWOG.

In Table 3, direct costs for Statistical Center personnel and operating expenses are estimated at $7304 per month. The largest single monthly cost is $3089 for psychologist and statistician effort. In Table 4, we estimate that every patient registered in a QOL study is associated with direct costs of $261. Table 4 also extends these cost estimates to the real world of total costs (i.e., [1.7 {direct costs}]). The total cost per patient on a QOL study is $443. An earlier examination of direct and indirect costs included only patients registered to trials coordinated by the SWOG. Under these restrictions, total costs were estimated to be $604 per patient registered and followed in a study with QOL assessments. As noted above, SWOG patients

Table 3. Estimated QOL personnel and operating costs/month*

Cost item		Cost per month
Personnel†		$5494
Operations office staff	(15%)	$ 516
Statisticians/psychologist	(48%)	$3089
Programmers	(17%)	$ 968
Data technicians	(24%)	$ 514
Data coordinator	(12%)	$ 407
Operating expenses‡		$1810
Direct costs per month		$7304

*Costs in 1994 dollars.

†Salaries include either a 22.5%, 25%, or 28% fringe benefit rate, depending on position.

‡QOL percentage (6%) of monthly operating costs for the Statistical Center includes secretarial and administrative staff salaries, two supply categories, postage, phone, and photocopying. QOL-related travel by the psychologist represents average monthly travel and does not involve the 6% calculation. There were 5174 phase 1, II, III, and other (e.g., cancer control) patient registrations for 1994, of which 225 were QOL registrations [(5174/12)/(331/12) = 0.06].

Table 4. Estimated total costs of QOL data per QOL patient* (averaged over life of current and closed protocols)

Direct QOL costs per month	$ 7304
No. of QOL registrations/month	28
Direct costs per QOL registration ($7304/28)	$ 261
Total QOL costs per month	$12 417†
Total costs per QOL registration (12 417/28)	$ 443

*Costs in 1994 dollars.

†Total cost = 1.7 (direct cost).

registered to QOL studies coordinated by other cooperative groups have been included because they do require registration and monitoring time on the part of some SWOG staff. However, the 106 1994 registrations to QOL studies coordinated by other groups involved less staff time than that required for the 225 patients registered to SWOG-coordinated protocols.

Discussion

Overestimate or Underestimate of Costs?

The goal of this project was to determine how much staff time currently was attributable to QOL tasks and to try to attach a dollar figure to that effort. The estimated direct and total costs for patients in QOL studies reflect the personnel-intensive nature of this research. The $443 per QOL registration cost is particularly interesting because the Statistical Center estimates that total cost (direct plus indirect) per patient in a therapeutic trial is very similar. The logs maintained by staff for 1 month reflect QOL tasks associated with both new QOL registrations, follow-up data, and data analysis at the close of studies. One could argue that $443 is an overestimate of the total costs, given that QOL registrations will fluctuate depending on the number of open studies. Consideration of patients registered in 1992 increases the QOL patient group from 331 to 423; the 1992 monthly average of 35 QOL patients per month reduces QOL total costs per patient to $355. However, in 1992, we had primarily newly activated protocols and no data cleaning or analysis activities. A log completed for 1 month during this period would have failed to include data analysis effort and would underestimate programmer time because the QOL monitoring system became more intensive beginning in the last half of 1993.

It should also be noted that costs displayed in Tables 3 and 4 are based on QOL studies that are companion or stand-alone studies leading to an underestimate of costs to do QOL research. That is, QOL companion studies require a separate therapeutic trial (with its associated costs) to provide the intervention generating the QOL hypotheses. One could estimate that the cost of a QOL study is $443 plus the cost of a therapeutic registration, for which we do not have detailed cost estimates; the total cost could be in excess of $1000. Since therapeutic trial data are being collected anyway, this is also an overestimate, but most QOL studies are dependent on the design and input of the therapeutic trial.

Although QOL studies will increasingly be integrated into therapeutic protocols (i.e., a single protocol including QOL end points), we do not expect integrated protocols to decrease QOL costs substantially. The real basis for decreasing cost will be moving beyond the substantial start-up costs associated with adding new data to the monitoring, database management, and analysis systems in place for clinical data. Our experience suggests that it takes 4-5 years to incorporate the QOL data into a cooperative group's clinical database and centralized monitoring scheme. At that point, programming tasks become somewhat more routine but certainly not eliminated. Methodologic work by the statisticians to deal with QOL analysis issues, particularly nonrandom missing data, still presents a substantial expenditure of effort; that is, data analysis is not yet routine.

The costs presented in Tables 3 and 4 underestimate monthly and per unit costs of collecting QOL data in SWOG trials in one very important area. Table 1 does not include the time spent by data managers at SWOG institutions establishing systems to ensure that the QOL assessment schedule is followed, collecting data from patients, reviewing data for completeness, and submitting questionnaires; institution staff must also see that patients whose clinical follow-up occurs at another site maintain the QOL assessment schedule. In addition, Tables 2-4 do not include the cost of QOL study coordinators who monitor questionnaire submission and, when possible, send reminders to institution staff regarding upcoming assessments. For three SWOG QOL studies, QOL coordinators reported that they spent 3 hours per month performing monitoring tasks; a fourth coordinator for a smaller, slow-accruing study spent 2 hours per month monitoring questionnaire submission. It would be informative to add the time spent by institution staff and then to model costs for all personnel using national salary data. This approach might yield more generalizable estimates. However, we do believe that the estimated level of effort and associated costs are reasonable central office estimates for the cooperative group context, with some variation due to regional differences in operating and personnel costs.

QOL Costs: Fixed Versus Variable?

One could argue that we are not really describing QOL costs per patient or we would define protocol development costs as a fixed cost. However, we find it difficult to define fixed costs (invariant components of cost not dependent on number of patients) versus variable costs (those that vary with the number of QOL registrations or protocols with QOL) for QOL research. In our context, most costs are variable and depend both on the number of proposed protocols and the number of QOL patient registrations to different trials. Increases in the number of protocols affect almost all staff because of protocol development and programming activities, whereas patient increases affect primarily data entry and monitoring costs. The only position that might be labeled a fixed cost is that of the psychologist, whose primary responsibility is QOL research but who also works on other cancer control research. Other staff responsibilities include a broader range of trial areas, both within cancer control and more broadly across all SWOG trials (e.g., data entry and programming). As QOL protocols increase in number and more patients are accrued to QOL studies, the Statistical Center has to address competing priorities for staff time.

Possibly a better summary variable would be QOL costs per protocol developed, but at this early stage of QOL research in the SWOG, it is difficult to trust the stability of the current number of protocols. What we are really interested in is QOL costs per a combination of QOL patients and protocols. The closest we come to describing that summary unit is the monthly cost data in Table 3. We are maintaining that costs attached to staff time during January 1995 reflect average monthly costs incurred by the SWOG attributable to QOL research. The costs per patient data in Table 4 are interesting but the preliminary conclusion may well be found in Table 3—the average cost per month of doing this research over the life cycles of SWOG trials that incorporate QOL assessments.

Volume of QOL data could also affect costs. The volume of QOL data is determined by questionnaire length and the number of times QOL is assessed in a trial. In the SWOG, questionnaire length varies primarily with respect to phase II and III trials. Phase II trials with a QOL component address a more restricted picture of QOL with a patient self-assessment of symptom status (usually 20 or fewer items); this assessment of a single QOL dimension is usually not referred to as QOL but patient report of symptoms. Phase III trials include a comprehensive patient self-assessment of QOL. The SWOG phase III questionnaire is longer (usually about 45-50 items) with additional subscales to measure physical, emotional, social, and role functioning as well as single-item measures of global QOL and comorbidity (*2*). The effect of this difference on central office costs is probably minimal, affecting primarily data entry and analysis time. The quality control and monitoring effort would not be differentially affected, since it is the submission of questionnaires, regardless of length, that drives the monitoring system (i.e., more like a fixed cost).

The second volume factor, number of assessment times, is more likely to affect costs, since the more times a QOL assessment is obtained, the greater the impact on all aspects of processing, particularly for quality control and monitoring (e.g., programming and staff monitoring time). The number of assessments in SWOG QOL studies has ranged from four times over several months (advanced-stage disease) to nine times over 7 years (early-stage disease) to once every 3-week treatment cycle (metastatic disease), which involved 15-24 assessments for some patients. The number of assessments is clearly a variable cost, which must vary with the course of the disease and the nature of the treatments under evaluation. Our staff log data were not detailed enough to address the impact of either type of volume on central office costs.

Funding Options

The SWOG initiated QOL research on a small scale, with the expectation that its mechanisms for collecting, monitoring, and processing clinical end point data could incorporate psychosocial data. Since the adoption of QOL assessment guidelines by the SWOG's Board of Governors in 1989, we have increasingly found it necessary to amplify quality control efforts. The increase has always involved more time and effort on the part of SWOG staff. If a cooperative group plans to use QOL data to help evaluate the overall effect of different cancer treatments, it must hold the collection, monitoring, and analysis of QOL data to the same strict standards in place for traditional medical end point data. Realistically, the extension of this commitment to new end point data requires additional resources.

Although QOL questions have received increasing support from clinicians engaged in clinical trials research, funding support for critical QOL data collection, quality control procedures, and analyses have not kept pace with the increased demand for inclusion of QOL measures in trials. One exception to this is NCI support for QOL methodologic research (*19*). Lack of funding poses a dilemma for clinical investigators and cooperative group research bases that consider including QOL end points in their clinical trials. The funding dilemmas discussed below, although illustrated with specific SWOG examples, are relevant to any cooperative group QOL effort, since all groups are funded through the same funding mechanism and all groups have access to similar external (to the cooperative group structure) funding options.

In the past, QOL staff time at SWOG institutions has primarily been funded through DCPC's cancer control credit program for Community Clinical Oncology Programs (CCOPs). Cancer control credits reimburse data manager and operational expenses for registering patients to cancer control protocols. QOL (and other cancer control) research at the Statistical Center and Operations Office has been funded through the SWOG's designation as a research base for CCOPs. As a research base, the SWOG develops cancer control protocols (including QOL protocols) and oversees data collection and analysis. At the inception of cooperative group QOL research, QOL studies were structured as companion or ancillary studies to therapeutic trials; companion studies were reviewed and approved solely by DCPC.

Currently, QOL studies are incorporated into the therapeutic protocol (*6*) and are subject primarily to DCT review. However, a protocol with a QOL component can be reviewed for cancer control credit by DCPC staff. Unless QOL end points are primary end points, award of cancer control credits requires an intervention other than the primary treatment arm evaluation (e.g., a symptom management or supportive care intervention). There is no funding mechanism in the DCT to cover the cost of QOL data collection, should the DCPC not approve a QOL study for cancer control credit. Furthermore, QOL issues are introduced not only by SWOG investigators but also by DCT staff in the protocol development phase. As noted above, the DCT has described guidelines for incorporating QOL end points in therapeutic trials (*6*).

SWOG will continue to apply for cancer control credits where QOL outcomes are considered primary and/or where they are linked to a cancer control intervention. However, even with the current credit system, the translation of a CCOP credit to funding for the Statistical Center results in considerably less than $443 total QOL costs per patient. In addition, the most cancer control credit awarded for a SWOG QOL companion study is 0.5 credit per patient registration; two studies were awarded 0.3 credits for each registration. This further widens the gap between Statistical Center costs and NCI reimbursement. Because there appears to be reduced support for QOL research through cancer control credits, SWOG has considered several cost reduction alternatives, some of which are more feasible than others for cooperative group research. 1) Request a line item in the DCT budget to cover the cost of adding QOL end points to therapeutic protocols. CALGB currently funds its telephone-based QOL data collection in this fashion. 2) R01 funding offers one funding option, but it is difficult to time funding requests and awards with the activation of a therapeutic protocol. This timing factor is important given the desire to begin QOL data collection with therapeutic trial activation. 3) Review requests for QOL studies proposed by SWOG investigators, selecting only a few to pursue. This review occurs at present but could be much more restrictive. 4) Try to reduce costs associated with how QOL data are handled. We, as have other cooperative groups, have become more efficient in processing QOL data and

expect some reductions as QOL monitoring systems become more institutionalized.

Conclusions

We have prepared these figures to generate renewed attention to QOL funding issues in cooperative group trials. They stem from our unwillingness to have less rigorous standards for QOL data than for clinical data and from our best estimate of the associated personnel and operating costs associated with expanding the scope of clinical trial end points. We will need to revisit this issue as we consider expanding the clinical trials database to include cost outcomes.

References

(*1*) Osoba D. Lessons learned from measuring health-related quality of life in oncology. J Clin Oncol 1994;12:608-16.

(*2*) Moinpour CM, Feigl P, Metch B, Hayden KA, Meyskens FL Jr, Crowley J. Quality of life end points in cancer clinical trials: review and recommendations. J Natl Cancer Inst 1989;81:485-95.

(*3*) Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality-of-life assessment in cancer clinical trials. Recent Results Cancer Res 1988;111:231-49.

(*4*) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international trials in oncology. J Natl Cancer Inst 1993;85:365-76.

(*5*) Nayfield SG, Ganz PA, Moinpour CM, Cella DF, Hailey BJ. Report from a National Cancer Institute (USA) workshop on quality of life assessment in cancer clinical trials. Qual Life Res 1992;1:203-10.

(*6*) Gotay CC, Korn EL, McCabe MS, Moore TD, Cheson BD. Quality-of-life assessment in cancer treatment protocols: research issues in protocol development. J Natl Cancer Inst 1992;84:575-9.

(*7*) Osoba D. The Quality of Life Committee of the Clinical Trials Group of the National Cancer Institute of Canada: organization and functions. Qual Life Res 1992;1:211-8.

(*8*) Ganz PA, Schag CA, Lee JJ, Sims MS. The CARES: a generic measure of health-related quality of life for patients with cancer. Qual Life Res 1992;1:19-29.

(*9*) Cella DF, Tulsky DS, Gray G, Sarafian B, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.

(*10*) Kaplan RM, Anderson JP. A general health policy model: update and applications. Health Serv Res 1988;23:203-35.

(*11*) Goldhirsch A, Gelber RD, Simes RJ, Glasziou P, Coates AS. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis [*see* comment citation in Medline]. J Clin Oncol 1989;7:36-44.

(*12*) Ganz PA, Lee JJ, Siau J. Quality of life assessment. An independent prognostic variable for survival in lung cancer. Cancer 1991;67:3131-5.

(*13*) Coates A, Gebski V, Signorini D, Murray P, McNeil D, Byrne M, et al. Prognostic value of quality-of-life scores during chemotherapy for advanced breast cancer. Australian New Zealand Breast Cancer Trials Group [*see* comment citation in Medline]. J Clin Oncol 1992;10:1833-8.

(*14*) Weeks J. Quality-of-life assessment: performance status upstaged? [editorial] [*see* comment citations in Medline]. J Clin Oncol 1992;10:1827-9.

(*15*) Maguire P. Using measures of psychological impact of disease to inform clinical practice. In: Ventafridda V, van Dam FS, Yancik R, et al., editors. Proceedings of the International Workshop on Quality of Life Assessment and Cancer. Amsterdam: Excerpta Medica, 1986:119-26.

(*16*) Ganz PA. Patient education as a moderator of psychological distress. J Psychosoc Oncol 1988;6:181-7.

(*17*) Skeel RT. Quality of life dimensions that are most important to cancer patients. Oncology 1993;7:55-61, 65-6, 69-70.

(*18*) Surveillance, Epidemiology, and End Results (SEER) Program/National Cancer Institute. Prostate cancer practice patterns and health-related quality of life. Special Study Statement of Work. N01-CN-05230.

(*19*) National Cancer Institute, National Center for Nursing Research. Quality of life assessment in special populations. RFA Number: CA/NR-92-27. In: Catalog of Federal Domestic Assistance No. 93.399, Cancer Control Research, and No. 93.361, Nursing Research, 1992.

(*20*) Shoemaker D, Burke G, Dorr A, et al. A regulatory perspective. In: Spilker B, editor. Quality of life assessments in clinical trials. New York: Raven Press, 1990:193-201.

(*21*) Aaronson NK, van Dam FS, Polak CE, et al. Prospects and problems in European psychosocial oncology: a survey of the EORTC Study Group on quality of life. J Psychosoc Oncol 1986;4:43-53.

(*22*) Aaronson NK. Quality of life assessment in clinical trials: methodologic issues. Controlled Clin Trials 1989;10:195S-203S.

(*23*) Sadura A, Pater J, Osoba D, Levine M, Palmer M, Bennett K. Quality-of-life assessment: patient compliance with questionnaire completion [*see* comment citation in Medline]. J Natl Cancer Inst 1992;84:1023-6.

(*24*) Kornblith AB, Anderson J, Cella DF, et al. Quality of life assessment of Hodgkin's disease survivors: a model for cooperative clinical trials. In: Tchekmedyian NS, Cella DF, editors. Quality of life in oncology practice and research. Williston Park (NY): Dominus, 1991:51-9.

(*25*) Hayden KA, Moinpour CM, Metch B, Feigl P, O'Bryan RM, Green S, et al. Pitfalls in quality-of-life assessment: lessons from a Southwest Oncology Group breast cancer clinical trial. Oncol Nurs Forum 1993;20:1415-9.

(*26*) Moinpour CM, Hayden KA, Thompson IM, et al. Quality of life assessment in Southwest Oncology Group trials. In: Tchekmedyian NS, Cella DF, editors. Quality of life in oncology practice and research. Williston Park (NY): Dominus, 1991:43-9.

(*27*) Moinpour CM, Savage M, Hayden KA, et al. Quality of life assessment in cancer clinical trials. In: Dimsdale JE, Baum A, editors. Quality of life in behavioral medicine research. Hillsdale (NJ): Lawrence Erlbaum Associates, 1995:79-95.

Notes

[1]Using an average working month of 173.3 hours possibly underestimates the proportion of staff time for QOL work since it does not account for vacation and sick leave that reduce the time actually spent on all work each month.

Supported by Public Health Service grants CA38926 and CA61674 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

I thank the following individuals for gathering information and for reviewing the manuscript: S. Condit, J. Crowley, S. Dahlberg, P. Feigl, M. Foster, M. Godfrey, G. Gullstrand, K. Hayden, L. Kingsbury, M. Lee, L. Loll, G. Lozano, B. McKnight, D. Marrah, E. Mize, K. Roth, M. Savage, S. Schmidt, J.M. Smith, J. Triplett, C. Upchurch, and N. Urban.

Modeling Health-Related Quality of Life: the Bridge Between Psychometric and Utility-Based Measures

*Pennifer Erickson**

Different purposes for assessing health-related quality, for example, clinical studies, epidemiologic analyses, and resource allocation, have led to the development and use of many different generic and disease-specific measures *(1-4)*. In addition to being categorized according to their purpose for development and application, measures can be grouped according to their conceptual frameworks, i.e., as being derived from either psychometric or utility theories of measurement. Two major distinctions between measures based on these conceptual frameworks are that 1) psychometric measures are essentially descriptive in nature whereas utility-based measures attempt to incorporate information about preferences for health states into the measure, and 2) utility-based measures can be combined with information on survival to incorporate mortality as well as morbidity into the assessment of health-related quality of life (QOL).

Need for a profile of scores, rather than a single summary number, as well as administrative constraints such as time and respondent burden, are reasons frequently given for selecting measures based on psychometric theory. At the same time, investigators readily acknowledge the desirability of having information from a utility-based measure, especially for evaluating gains in both quality and quantity of life associated with incremental costs of treatment and for understanding contradictory findings that may occur with a profile of scores rather than an overall summary score (*5*). Thus, although having data from both types of measures might be considered ideal, practical considerations usually result in the use of only one type of measure, depending on the purpose of the study.

Previous research has illustrated how data collected in descriptive studies could be transformed into a measure that has properties of a utility-based measure (*6*). In this paper, the earlier model is expanded to include both conceptual and statistical methods for generating and validating a measure that is developed from existing data. With this model, data that have been collected using the advantages of the psychometric method can be transformed to gain the analytic potential of a utility-based measure. The model is tested using two generic health-related QOL instruments and data from the 1987 National Medical Expenditure Survey (*7*). Some practical implications of modeling health-related QOL as a bridge between psychometric and utility-based measurements are discussed.

Conceptual Frameworks

Essential distinctions between psychometric and utility-based measures that are relevant for modeling are summarized in Table 1. Most measures that are based on a psychometric framework have standardized questionnaires that have been developed through a series of pilot tests. These tests are designed to identify appropriate concepts and domains for target populations, to correct ambiguously worded questions, and to select a reasonable minimum number of questions needed to measure a given concept. A part of the pilot testing of each data-collection instrument is also to determine its reliability and validity in various populations. The result is a questionnaire that can be used to make scientific inference about health-related QOL and that minimizes administrative and analytic burden for both the respondent and the investigator. Part of the ease of administration is due to the use of Likert scaling or other descriptive response categories that are generally easy for respondents to understand and for data analysts to edit and score. Subscale scores can be arrayed as profiles or polar graphs for ease of interpretation; graphic depictions assist the decision maker in forming an overall assessment of health status in the absence of a summary score. Among some of the frequently used measures that are founded in psychometric theory are the Short Form 36, the Functional Assessment of Cancer Therapy Quality of Life Questionnaire, the Functional Living Index—Cancer, and the Cancer Rehabilitation Evaluation System (*8-11*).

Utility-based measures that are most amenable to modeling health-related QOL are characterized as having a classification system. Utility-based measures that consist of a small number of health or disease-specific health scenarios for which utility weights are assigned directly using a scaling method, such as the standard gamble or time tradeoff technique, may be useful but they have not been considered in this model. The classification system, which may or may not have a standardized questionnaire, is used to categorize individuals into mutually exclusive health states that may be defined in terms of single concepts or attributes, e.g., activity limitation or perceived health, or holistically. Similar to the information obtained from a psychometric measure, the classification system provides descriptive information about the health of the study group. With an increasing number of concepts and domains being used to develop operational definitions of health, the single- and multi-attribute approaches are becoming the more widely used format.

**Correspondence to:* Pennifer Erickson, Clearinghouse on Health Indexes, Office of Analysis, Epidemiology, and Health Promotion, National Center for Health Statistics, 6535 Belcrest Rd., Rm. 730, Hyattsville, MD 20782.

Table 1. Distinguishing characteristics of psychometric and utility-based measures relevant for modeling health-related QOL

Characteristic	Measure: Psychometric	Measure: Utility-based
Conceptual frameworks	May include multiple concepts and domains of health-related quality of life	May include multiple concepts and domains of health-related quality of life
Standardized questionnaire	Standardized instrument available, usually with demonstrated measurement properties, such as reliability and validity	Standardized instrument may not be available
Items or function levels	Responses to items are descriptive categorical ratings, frequently scaled using Likert scaling	Responses are differentially weighted to reflect preferences for health states using rating scale, time tradeoff, or standard gamble methodology
Subscales representing concepts and domains	Scores for subdomains are formed by adding item scores; items are assumed to be equally important	Subscale scores are rarely presented
Overall score	May have an overall score; subscale scores are usually presented as a profile of scores	Overall score is calculated using multiattribute utility or holistic scaling methods

For utility-based measures, the health states in the classification system are differentially weighted according to the value or utility placed on the level of functioning shown in the state. Weights may either be taken from an existing set of utilities or be generated for a specific application following procedures associated with one of the accepted utility-elicitation schemes, usually a rating scale, time tradeoff, or standard gamble technique (*12*). These weights are used to form a summary score, or index, that represents the level of health of an individual. The weights may also be combined with survival information to express health-related QOL in terms of quality-adjusted life years or years of healthy life. Among some of the frequently used utility-based measures are the Health Utilities Index, Healthy People 2000 Years of Healthy Life, Quality of Well-Being Scale, and the Q-TWiST (*13-16*).

Both summary scores and profiles can be used to show relative increments or decrements within the same person over the course of treatment or between groups of patients with different treatments or different diseases. The differences between scores and profiles, in addition to those associated with ease of administration as discussed above, have to do with interpretability. Profiles, on the one hand, present scores for each of the different attributes measured in the profile; summarization of this information is done by individual decision makers. Indexes, on the other hand, present a summary score that is done consistently across different decision makers, but the overall score may obscure areas of individual dysfunction that need improvement. Thus, the ideal health-related QOL measure is thought to be one that gives an overall summary score and yet can be disaggregated to identify functional areas that might be targeted for treatment. The following model is suggested as a way of arriving at this ideal.

Model for Developing Utility-Based Measures

The model describes methods and rationale for transforming data collected using a psychometric framework, referred to here as the source of the data, into a classification system that is associated with the targeted utility-based measure. The development of this model builds on research that was designed to convert data collected by means of batteries of questionnaires into utility-based measures of health-related QOL (*6,17,18*). This earlier work represented a generic approach to retrospective analysis; that is, the model was not restricted to having the source and target datasets based on psychometric and utility-based measures, respectively.

The current adaptation of this generic approach uses both conceptual and statistical modeling to bridge between psychometric and utility-based measures of health-related QOL. The goal of conceptual modeling is to align the psychometric source data and the target utility-based measure so that they contain the same concepts and domains of health-related QOL to the extent possible. As indicated in Table 2, the first step is to critically and carefully review the questionnaire that was used to collect the data. In conducting this review, analysts identify aspects of the questionnaire and the data-collection process, such as question framing and recall period, that are likely to influence responses about the type and degree of dysfunction reported.

With this comprehensive understanding of the source data, the next step is to review existing utility-based assessments to identify the one that is most like the source in terms of health-related QOL content. For both the Health Utilities Index Mark I and the Quality of Well-Being Scale, the review indicated by Step 2, Table 2 has been completed (*6,19*). In certain situations, however, it might be desirable to enhance these existent reviews if additional information is needed.

Once the utility-based measure has been identified, items from the source questionnaire are matched according to the concept of health-related QOL and question-design issues with health states in the utility-based measure. Items representing comparable levels of function are identified and used to develop an analogue of the classification system. During this process, concepts and questionnaire design features that differ between the two might be identified. In most studies, the differences will occur because of data limitations. In some studies, however, these differences may occur by design. For example, the analyst may choose to use only a subset of the concepts included in the utility-based measure as relevant in the current study.

Statistical modeling is done after a classification system has been constructed from items in the psychometrically based

Table 2. Modeling health-related quality of life: steps for converting data collected according to a psychometric measurement strategy into a utility-based measure

Conceptual modeling
Review the questionnaire used to collect data, i.e., the source of the information in terms of concepts and domains of health-related QOL included in the questionnaire:
Question framing, e.g., performance or capacity mode
Recall period, e.g., 1 day, 2 weeks, 1 month
Respondent, e.g., self or proxy
Evaluate and select a utility-based measure to serve as the target, based on the following criteria:
Concepts and domains, question framing, recall period, and respondent that are similar to those in the source
Minimal discrepancies between possible target classification systems and the source of data
Construct an analogue of the target classification system using the psychometric source:
Include all items from the source questionnaire that are used in developing the utility-based measure in the corresponding "cells" of the classification system
List all assumptions necessary to convert the source into the target classification system
Specify decision rules for handling missing data
Statistical modeling
Test the content and face validity of the constructed classification system by using:
Criterion-type validity if external data sources are available
Regression modeling to test for relationships
Construct and validate scores using:
Scoring algorithm specified for the utility-based measure
Construct validity of the overall scores
Conduct regression as well as descriptive analyses to determine how the constructed scores compare with known health status and quality-of-life relationships
Conduct sensitivity analysis to:
Identify the impact of the assumptions and decision rules
Assess the degree of confidence that can be placed on inferences drawn from use of the measure

questionnaire. Content and face validity can be assessed using descriptive analyses to examine response patterns within and between known groups. Criterion-type validity can be assessed by comparing prevalence of dysfunction observed with the constructed health states with prevalence of the same dysfunction observed in an external data source. For example, in validating a Health Utilities Index Mark I classification system that was constructed using data collected in the NHANES I Epidemiologic Followup Study, data from the National Health Interview Survey were used to compare estimated percentages of dysfunction in activities of daily living and other forms of physical and role limitations (*17*).

After criteria for content and criterion-type validity of the classification system have been met, utility weights can be assigned and the overall scores computed for each individual in the study group. Convergent construct validity can be assessed by forming various hypotheses about the relationships of scores between known groups and between groups of persons defined in terms of their health characteristics and health-care use. In addition, regression analyses might be conducted to determine the impact of various diseases and utilization patterns on health when other personal and lifestyle characteristics are held constant.

The final step is to use sensitivity analysis to estimate the effect of the assumptions that were made during the process of constructing the classification system or assigning the scores. Varying the assumptions to give a range of scores indicates the robustness of the constructed measure. If changing the assumptions has little or no impact on the constructed measure, then this increases the degree of confidence that can be placed on the scientific inferences that might be made when using the constructed utility-based measure.

Application of This Model

This model has been used to convert data collected as part of the National Medical Expenditure Survey (NMES) into scores that are based on the Health Utilities Index Mark I. The NMES is a national panel survey that was conducted in 1987 (*6*). Individuals were selected to participate in NMES using a complex sampling procedure so that the resulting sample is representative of the U.S. population. Data from approximately 20 000 adults who completed the self-administered Health Status Questionnaire that was administered via a postal survey in the spring of 1987 have been used to develop scales that are comparable to the Physical Function, Mental Function, and General Health Perceptions scales of the Medical Outcomes Study Short Form (*18,20,21*).

Mean scores for Physical Function and General Health Perceptions decline with age, whereas the scores for Mental Function are relatively constant across age (Table 3). For all scales, males have higher scores than do females, and whites have higher scores than do blacks. Data are shown for persons who report that they have or do not have arthritis to illustrate the sensitivity of these scales to the impact of a chronic disease that affects QOL but has little or no impact on quantity of life. As might be expected, mean scores are lower for persons with arthritis for all age groups and for all three subscales.

To model the Health Utilities Index Mark I (HUI-I), the goal was to match information from the NMES version of the Medical Outcome Study Short Form with the original HUI-I classification system that includes four major domains: 1) Physical Function: Mobility and Physical Activity; 2) Role Function: Self-Care and Role Activity; 3) Social–Emotional Function: Emotional Well-Being and Social Activity; and 4) Health Problems (*22,23*). Each domain is described in terms of a set of mutually exclusive levels. To model an analogue of the HUI-I, the composite functions in the original classification system were disaggregated into 23 levels that each represented one type of function. The goal was to find information in the NMES Health Status Questionnaire that indicated whether each survey respondent did or did not have the dysfunction indicated in each of the disaggregated levels.

Some levels of the Health Utilities Index Mark I, e.g., those in the Health Problems domain, are not included in the Medical Outcomes Study profile but were asked as part of the NMES Health Status Questionnaire. These additional items were used to construct a more complete HUI-I classification system than would have been possible from the Medical Outcomes Study alone. Once all of the matches were made and the validity of the constructed classification scheme assessed by comparing prevalences of dysfunctions with those obtained in either the National Health Interview Survey or the NHANES I Epidemiologic Fol-

Table 3. Means and SEs for Physical Function, Mental Function, and General Health Perceptions Subscales by age group for selected demographic and health characteristics, National Medical Expenditures Survey, 1987

Population group	Total		18-34 y		35-54 y		≥55 y	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Physical function								
Male	87.42	0.29	96.10	0.25	91.07	0.39	71.41	0.70
Female	82.55	0.32	94.25	0.30	86.87	0.42	65.46	0.58
White	85.02	0.28	95.39	0.22	89.46	0.33	68.92	0.56
Black	81.92	0.71	94.23	0.68	84.55	0.89	58.28	1.29
With arthritis	62.27	0.64	82.85	1.67	71.55	0.86	56.41	0.66
Without arthritis	91.15	0.18	95.72	0.20	92.48	0.25	76.19	0.51
Total	84.86	0.27	95.16	0.21	88.94	0.29	68.07	0.54
Mental function								
Male	76.03	0.24	76.87	0.29	75.83	0.42	75.18	0.44
Female	71.89	0.23	72.13	0.36	72.14	0.33	71.37	0.32
White	74.03	0.22	74.34	0.29	74.21	0.33	73.49	0.36
Black	71.99	0.46	74.13	0.64	72.05	0.62	68.27	0.84
With arthritis	68.01	0.36	67.50	1.18	66.22	0.69	68.78	0.41
Without arthritis	75.50	0.20	74.73	0.26	75.54	0.27	77.10	0.39
Total	73.83	0.21	74.41	0.26	73.94	0.30	73.03	0.34
General health								
Male	68.11	0.43	76.59	0.44	71.09	0.63	52.94	0.74
Female	65.40	0.36	73.76	0.40	68.55	0.57	52.57	0.59
White	67.40	0.37	76.02	0.34	71.07	0.53	53.59	0.60
Black	60.61	0.66	70.23	0.73	61.25	0.93	42.51	1.06
With arthritis	47.71	0.54	60.34	1.57	53.87	1.04	43.72	0.65
Without arthritis	71.95	0.30	75.81	0.32	73.06	0.46	61.17	0.58
Total	66.68	0.35	75.12	0.31	69.79	0.49	52.73	0.56

lowup Study, overall HUI-I scores were assigned to each individual in the survey according to standard scoring procedures (*22*).

In matching the NMES data with the HUI-I classification system, the three following situations occurred. One was that the data collected in NMES were a close match with the conceptual content of the HUI-I; for example, both NMES and HUI-I include information about limitation in ability to bend. Another was that the NMES data were a likely, but less than perfect, match with the content of the HUI-I; for example, the NMES asks about trouble walking one block, while the HUI-I classifies people according to limitation in physical ability to walk without specifying a distance. The third situation was when there was no clear match between the two. For example, the HUI-I classifies people according to ability to run or jump; the closest NMES item to this concept is the kind or amount of vigorous activities that the respondent can do. When there was less than a close match, information was used when it was possible to classify persons without introducing significant bias; the assumptions were carefully noted.

The patterns of mean scores for the constructed NMES-HUI-I score (Table 4) are essentially the same as those for Physical Function, Mental Function, and General Health Perceptions Subscales (Table 3). Males have higher mean scores than do females; the white population has higher mean scores than does the black population; and the persons with arthritis have lower mean scores than do those without arthritis. In addition, mean scores are highest for persons in the youngest age group and lowest for persons in higher age groups. These patterns, as well as more detailed comparative analyses, indicate that the NMES-HUI-I is a valid measure of health-related quality of life.

This application of the NMES data to the model shown in Table 2 indicates that valid utility-based measures can be developed from data that have been collected using a questionnaire that is based on measurement principles that have been

Table 4. Means and SEs for the National Medical Expenditures Survey-HUI-I by age group for selected demographic and health characteristics—1987

Population group	Total		18-34 y		35-54 y		≥55 y	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Male	0.84	0.00	0.89	0.00	0.87	0.00	0.74	0.01
Female	0.80	0.00	0.86	0.00	0.83	0.00	0.70	0.01
White	0.82	0.00	0.88	0.00	0.85	0.00	0.73	0.00
Black	0.79	0.01	0.88	0.01	0.82	0.01	0.63	0.01
With arthritis	0.66	0.00	0.75	0.01	0.70	0.01	0.63	0.01
Without arthritis	0.87	0.00	0.88	0.00	0.88	0.00	0.80	0.00
Total	0.82	0.00	0.88	0.00	0.85	0.00	0.72	0.00

derived from psychometric theory. Practical implications of modeling health-related QOL are discussed below.

Discussion

This model is intended to serve as a bridge between data collected using one conceptual framework and analyses using an alternative framework and thus has practical implications for conducting research on health-related QOL. One implication is that existing descriptive data on health-related QOL can be reanalyzed as a utility-based measure without additional data collection. From a clinical trial perspective, such a reanalysis might be desirable if data from the health-related quality-of-life profile or battery of measures are giving contradictory findings. For example, over the course of the study, some of the concepts of health measured in the profile may be showing increments in health status, whereas others may be showing decrements. By converting various concepts and domains of health-related QOL into a summary score, the overall net effect of the trial, whether a net increase or decrease in QOL, might be more readily apparent.

For long-term trials in which mortality is an important health outcome, the ability to convert data from a battery or profile of scores into a measure that allows for the inclusion of death may be important in determining the full impact of the treatment regimens, that is, not only the impact on QOL but also on quantity of life. Similarly, modeling health-related QOL may be useful for epidemiologic analyses that examine determinants of health of cohorts of individuals across time.

Third, retrospective analyses of existent data can also be useful in situations when it is desirable to use data collected as part of a clinical research protocol to obtain some indication of the cost implications of a new treatment. Since utility-based measures can be combined with mortality data to indicate years of healthy life, they have been recommended for use in cost–utility analysis, a variant of cost-effectiveness analysis. In addition, the years of healthy life metric is readily understood by many people, since it converts information on QOL into a biologically meaningful outcome measure. Thus, modeling can be used to expand the potential usefulness of the data from descriptive portrayals of the study results to more analytic interpretations of the findings.

The model or a similar type of mapping strategy might be actively considered when designing a prospective study, whether a clinical trial or a cost-effectiveness analysis. Data can be collected using standardized, reliable, and valid questionnaires that minimize respondent and analytic burden. These descriptive data can subsequently be combined with existing utility weights, thereby eliminating the need to conduct study-specific utility or preference elicitation studies that can be very costly to administer and interpret. Thus, although the illustration of this model through the use of the NMES dataset might be interpreted by some as indicating that it is only useful when the investigator lacked the foresight to collect the desired data, the potential for analyzing data collected by means of a psychometric format according to utility-based measurement models might be actively considered by investigators as a strategy for efficiency in study design.

When constructing a utility-based measure from data collected using either a health profile or battery of measures, the modeled measure is not strictly comparable to the original. As noted in the discussion of the model (Table 2), sensitivity analysis indicates some of the extent to which the original and constructed measures differ. Conservative use of the constructed measure restricts inferences based on the modeled measure to the database that served as the basis for the constructed measure; if other databases have been developed using the same survey procedures and instruments, these may also be used for comparing treatments and drawing inferences. Comparisons of results using an original with a constructed measure are subject to the same restrictions as are any attempts to draw inferences across databases that have been developed using different methods.

References

(*1*) McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. New York: Oxford University Press, 1987.

(*2*) Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989;27:S217-32.

(*3*) Patrick DL, Erickson P. Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford Univ Press, 1993.

(*4*) Berzon RA, Simeon GP, Simpson RL Jr, Donnelly MA, Tilson HH. Quality of life bibliography and indexes: 1993 update. Qual Life Res 1995;4:53-74.

(*5*) Ganz PA. Quality of life and the patient with cancer. Individual and policy implications. Cancer 1994;74:1445-52.

(*6*) Erickson P, Kendall EA, Anderson JP, Kaplan RM. Using composite health status measures to assess the nation's health. Med Care 1989;27(3 Suppl):S66-76.

(*7*) Edwards WS, Berlin M. National Medical Expenditure Survey methods 2: questionnaires and data collection methods for the household survey and the survey of American Indians and Alaska Natives. Rockville (MD): Agency for Health Care Policy and Research: DHHS Publ No. (PHS) 89-3450.

(*8*) Schipper H, Clinch J, McMurray A, Levitt M. Measuring the quality of life of cancer patients: the Functional Living Index—Cancer: development and validation. J Clin Oncol 1984;2:472-83.

(*9*) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.

(*10*) Ganz PA, Schag CA, Cheng HL. Assessing the quality of life—a study in newly-diagnosed breast cancer patients. J Clin Epidemiol 1990;43:75-86.

(*11*) McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-item Short-Form Health Status Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247-63.

(*12*) Torrance GW. The measurement of health state utilities for economic appraisal. J Health Econ 1986;5:1-30.

(*13*) Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index—Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995;4:249-57.

(*14*) Erickson P, Wilson RW, Shannon I. Years of healthy life: statistical Note 7. Hyattsville (MD): National Center for Health Statistics, 1995.

(*15*) Kaplan RM, Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. Health Psych 1982;1:61-80.

(*16*) Gelber RD, Goldhirsch A. A new endpoint for the assessment of adjuvant therapy in postmenopausal women with operable breast cancer. J Clin Oncol 1986;4:1772-9.

(*17*) Erickson P, Kendall EA, Odle MP, Torrance GW. Assessing health-related quality of life in the National Health and Nutrition Examination Survey. Hyattsville (MD): National Center for Health Statistics, 1993.

(*18*) RAND Health Sciences Program. RAND 36-item health survey 1.0. Santa Monica (CA): The Rand Corporation, 1992.

(*19*) Erickson P, Anderson JP, Kendall EA, et al. Using retrospective data for measuring quality of life: National Health Interview Survey data and the Quality of Well-being scale. Qual Life Cardiovasc Care 1988;4:179-84.

(20) Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, et al. Functional status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. JAMA 1989;262:907-13.

(21) Stewart AL, Ware JE Jr, editors. Measuring functioning and well-being: the Medical Outcomes Study approach. Durham (NC): Duke Univ Press, 1992.

(22) Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Univ Press, 1987.

(23) Torrance GW. Multiattribute theory as a method of measuring social preferences for health states in long-term care. In: Kane RL, Kane RA, editors. Values and long-term care. Lexington (MA): D.C. Heath and Company, 1982:127-56.

Taking Quality of Life Into Account in Health Economic Analyses

*Jane Weeks**

Cost–utility analysis is the most commonly used approach to incorporating quality-of-life considerations into economic analyses in health care. This type of analysis produces a ratio of the incremental cost of one intervention over another to the incremental benefit produced, measured in quality-adjusted life years. To be suitable for use in calculating quality-adjusted survival, quality of life must be measured in the form of a utility. Direct utility assessment techniques are grounded in decision analytic theory and are conceptually complex and impractical for use in the clinical trial setting. Alternatives include global rating scale items with appropriate "transformations" and health state classification indices. The first cancer trials to collect economic data and utilities from patients using these techniques are now under way. These trials will serve to answer not only biological questions, but also health policy questions about whether the additional cost of the more expensive therapy is justified by the benefit it produces in both length and quality of life. [Monogr Natl Cancer Inst 1996;20:23-7]

The goal of any health economic analysis is to determine whether the cost of a particular intervention is justified by the health benefits it produces. The question is usually framed by asking not how much it costs to deliver a particular treatment, but how much *more* it costs to provide that treatment than the most reasonable alternative (*1*). This alternative may be a "no-treatment" strategy, but this does not necessarily mean it is a "no-cost" strategy.

All economic analyses examine the difference in cost between alternative strategies; they differ in how they measure the benefits resulting from those strategies (*1,2*) The four basic types of economic analysis measure these benefits in four different ways.

A *cost-minimization* study simply assesses the additional cost of one strategy in comparison with another and therefore implicitly assumes that the two treatments produce comparable benefits. Because alternative medical interventions rarely produce truly equivalent outcomes, this type of analysis generally does not suffice as a complete economic evaluation of competing interventions. Usually, one wants to know whether the additional benefit conferred by the more expensive treatment is sufficient to justify the additional cost.

Cost-benefit analyses answer this question by assigning a dollar value to the health outcome in order to determine whether the incremental benefit of one treatment over another, measured in monetary terms, is greater than or equal to the incremental cost.

Cost-effectiveness analyses, in contrast, measure the benefits of health care interventions in units of medical effect. For example, the cost-effectiveness of combination chemotherapy compared with single-agent therapy for a given disease could be assessed by calculating the additional cost (in dollars) per additional patient reaching the 5-year disease-free survival mark. One of the goals of cost-effectiveness analysis, however, is to facilitate resource allocation decisions between interventions to treat or prevent different diseases. Cost-effectiveness data are much more useful if health benefits are measured in units that are common across diseases. The most frequently used measure is years of life saved. Cost–effectiveness ratios are therefore usually expressed in terms of dollars per year of life saved.

But medical interventions affect not only length of life but also quality of life. Cancer cure may be bought at the expense of substantial treatment-related morbidity. Conversely, palliative therapy may bring marked relief of symptoms even if it does not lengthen life dramatically. *Cost–utility* analysis, a specific type of cost-effectiveness analysis, takes into account the impact of a health intervention on quality of life as well as length of life. Most commonly, this is done by assessing health benefits in terms of quality-adjusted survival, measured in quality-adjusted life years (QALYs). The units of a cost–utility ratio are thus dollars per QALY.

Approaches to Measuring Quality of Life for Economic Analysis

In the 48 years since Karnofsky et al. (*3*) initiated the measurement of health status in cancer patients, a number of sophisticated instruments have been designed that assess cancer patients' health-related quality of life (HRQOL) in multiple dimensions. About the same time that Karnofsky et al. first assessed functional status in cancer patients, von Neumann and Morgenstern (*4*) developed the foundations of assessing utilities, defined as strengths of preferences for various health states. HRQOL research thus evolved out of at least two theoretical traditions. The legacy of this historical development is two over-

**Affiliation of author:* Center for Outcomes and Policy Research, Division of Cancer Epidemiology and Control, Dana-Farber Cancer Institute, Boston, MA.

Correspondence to: Jane Weeks, M.D., M.Sc., Center for Outcomes and Policy Research, Division of Cancer Epidemiology and Control, Dana-Farber Cancer Institute, 44 Binney St., Boston, MA 02115.

See "Note" section following "References."

lapping but distinct approaches to the measurement of HRQOL; one approach is based on measures of health status, and the other is based on measures of preferences or utilities.

Health status measures collect information on physical and psychosocial functioning, usually in a number of domains or dimensions, including physical functioning, mood, social support, and health perception. A number of instruments to measure these dimensions have been developed and have been shown to be reliable and valid in diverse populations of cancer patients (*5-11*). They have proven to be effective tools in generating descriptive data on the experience of cancer patients with different stages of disease (*6*) and have been applied successfully to compare outcomes in groups with relatively stable clinical states, such as survival after childhood cancer or localized breast cancer (*12,13*).

These scales are less useful, however, in comparing alternative treatment strategies that result in time-dependent changes in health status, in assessing the appropriateness of trade offs between quality and quantity of survival, or in determining whether the benefits of medical therapy justify the costs. The ideal HRQOL measures for this purpose involve measuring the value of health states by reference to a universal standard such as time, money, or risk of death. Such measures are called "utilities." The terms "values" and "preferences" are often used as synonyms. By convention, utilities are measured on a scale of 0-1; 0 represents death, and 1 represents excellent health.

Utilities differ from more familiar measures of quality of life in that they reflect how a patient *values* a state of health, not just the characteristics of the health state. They are an appealing measure of global HRQOL because respondents rather than researchers determine the importance or weight to assign to each domain in calculating overall HRQOL. More importantly, because of the way these questions are structured, the utilities they generate can be multiplied by the length of time spent in that health state to produce a single measure that reflects both quality and length of life. Therefore, unlike health status measures, utilities can be used to calculate quality-adjusted life survival, which reflects the area under an HRQOL versus time curve (*14*). Data on the quality-adjusted survival resulting from alternative treatment strategies have two major uses. First, this information may help patients and their physicians assess the trade offs between length and quality of life inherent in many decisions about cancer therapy. In particular, for the patient who is overwhelmed by a presentation of comprehensive data on the survival and quality-of-life outcomes of alternative treatment strategies, it may be very useful to know which alternative produces the best quality-adjusted survival for the "typical patient" (*15*). Second, quality-adjusted survival provides a useful measure of the benefit of medical therapies for public policy discussions and decisions. It permits comparisons of the value of health care interventions across diseases and is the standard measure of benefit in cost-effectiveness analyses.

Techniques of Utility Measurement

A respondent's utility for a given health state may be elicited in several different ways. The simplest approach is to use a rating scale. The basic structure of a rating scale is that it is a continuous measure anchored by descriptors at both ends. The standard descriptors in a global rating scale of overall health-related quality of life are "excellent health" and "death." The rating scale may be presented in the form of a visual analog scale, "feeling thermometer," or a verbal numeric scale.

The big advantage of a rating scale is that it can be easily self-administered. Unfortunately, it does not produce a true utility. There is no reason to believe that a respondent who assigns a state of health a score of 75 on a 100-point rating scale would be willing to give up exactly one quarter of his or her life expectancy in exchange for a return to perfect health.

True utility measures can be interpreted in this fashion, however, because they ask about quality of life in exactly these terms. The classical utility measure is the standard (or reference) gamble (*16*). This technique assesses a respondent's utility for his or her own quality of life (or that of a hypothetical health state) by asking how much risk of death he or she would accept to improve quality of life. In a standard gamble, the respondent is asked to choose between life in a particular health state with less than perfect quality of life and a gamble between death and perfect health. The probability of death in the gamble is systematically varied until the respondent is indifferent between the gamble and the certain, intermediate outcome. The respondent's utility for the health state is given by the probability of perfect health in the gamble at which this point of indifference is reached. One salient feature of the standard gamble is that the elicited utility reflects not only the respondent's preferences about the quality of life in the health state but also whether he or she is a risk taker or a gambler.

An alternative utility measure that is not influenced by the respondent's attitude toward risk is the time trade-off. This technique assesses the respondent's utility for a health state by asking how much time he or she would give up to improve it. The respondent is offered a choice between a set length of life in a given compromised health state and a shorter length of life in perfect health. The respondent's utility or strength of his or her preference for the compromised health state is given by the ratio of the shorter to the longer life expectancy at which the responent finds the two choices equally desirable.

Both the standard gamble and the time trade off are conceptually complex. They require the respondent to grasp hypothetical scenarios, to manipulate probabilities and life expectancies, and to confront the possibility of imminent death. Anecdotal evidence suggests that respondents who are older or less educated have particular difficulty comprehending these items. Comprehension may be improved by administering the questions in an in-person interview using visual aids to demonstrate the probabilities involved or by computer programs designed specifically for this purpose (*17,18*). But these techniques are not well suited for use in the clinical trial setting; as a result, standard gambles and time trade offs are almost never used to collect utilities in clinical trials.

Therefore, there is great interest in alternative approaches to utility assessment that can be self-administered in a paper-and-pencil format. Rating scales are often used in this fashion even though they are not true utility measures. Some studies (*19,20*) have demonstrated that the mean utility for a population is reasonably

well correlated with the mean rating scale value if that value is "transformed" to adjust the score upward. For example,

$$\text{utility} = 1.18 \times (\text{rating scale}), \text{ for rating scale} < 0.85,$$
$$\text{and utility} = 1, \text{ for rating scale} \geq 0.85.$$

Other appealing alternatives to direct utility assessment are "hybrid" approaches that maintain the ease of administration of a traditional quality-of-life questionnaire, while also producing utility estimates appropriate for use in clinical and economic decision making. These health state classification indices consist of two components: 1) a simple health-related quality-of-life questionnaire that is completed by patients to generate descriptive data and 2) a formula that assigns a utility to each patient's set of responses to that questionnaire (Fig. 1). The formula reflects the relative importance or weight assigned to different domains of health-related quality of life by respondents in a reference population. Examples of such systems include the Quality of Well-Being Index (*21*) and the Health Utility Index (*22*). Approaches currently undergoing validation include EuroQol (*23*), a measure specifically designed for international use, and the Q-tility Index (*24*), a cancer-specific tool.

What is the justification for turning to a reference population rather than patients themselves for the preference weights for such a system? It is commonly argued that, for purposes of health policy decisions, it is appropriate to use a general population reference group, since society's preferences should determine how society's resources are allocated. A case can also be made that the relevant preferences for medical decision making are those of a respondent evaluating an array of potential outcomes rather than those of a patient experiencing one particular health outcome. Health state classification indices therefore rely on patients to provide information on the nature of the impact of a given health state on quality of life but use proxy decision makers as the source of the weights for generating a utility score for that state.

Estimating Cost–Utility Ratios

The most common approach to estimating the incremental cost–utility of one medical intervention in comparison to another is to rely on decision-analytic modeling to estimate

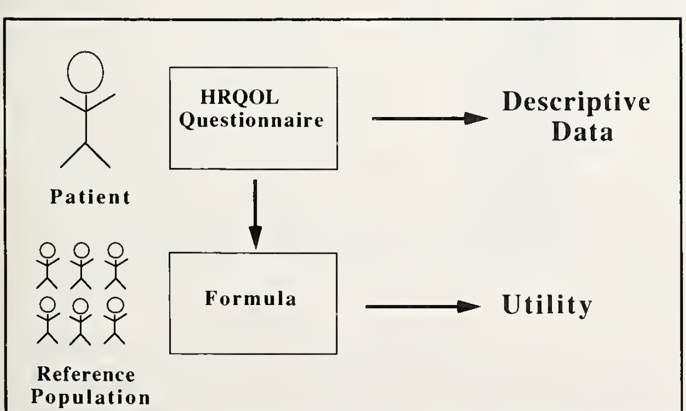

Fig. 1. Components of a health state classification index. HRQOL = health-related quality of life

quality-adjusted survival. In such models, health state outcomes are assigned utility values in a decision tree or Markov model. These models use data on the probability of various outcomes to generate estimates of quality-adjusted survival expected from the interventions considered in the model. Until recently, the utility estimates used in these models were nearly always based on "expert opinion" (e.g., guesses by the modeler and perhaps a few colleagues). Increasingly, polls of health professionals or focus groups with patients serve as the source of the utility elements in these models.

In recent years, investigators have begun to turn to clinical trials instead as the source of all data needed to perform cost–utility analyses, including not only biologic outcomes but also economic and utility data. The first U.S. cancer cooperative group trial to include a prospective cost-effectiveness analysis serves as one example of how this might be done. Intergroup Trial 0146 ("A Phase III Prospective Randomized Trial Comparing Laparoscopic-Assisted Colectomy Versus Open Colectomy for Colon Cancer" [Principal Investigator: Heidi Nelson]) is one of several studies being funded by the U.S. National Cancer Institute (NCI) in response to a Request for Applications on minimal access surgery in cancer treatment. In an unprecedented move, the NCI required that all studies submitted for consideration for funding through this mechanism include evaluations of economic and quality-of-life outcomes.

This study, which began accrual in late 1994, randomly assigns patients with newly diagnosed colorectal cancer to receive either laparoscopic-assisted or open colectomy (*25*). The primary end point for the study is cancer recurrence. Quality of life, cost, and cost–utility are secondary end points.

The quality-of-life component of the trial includes evaluation of symptoms as well as quality of life per se. Patient self-reported symptoms are assessed using the Symptom Distress Scale (*26*) completed at study entry and 48 hours, 14 days, and 2 months after surgery. Quality of life is measured with the Quality of Life Index (*7*) at study entry and 14 days, 2 months, and 18 months after surgery. Utilities are assessed at these same time points using a rating scale of 0-100 of overall quality of life and the Q-tility Index (*24*), a cancer-specific health state classification system that assigns a utility to any set of responses to the Quality of Life Index. This combination of instruments was selected to maximize responsiveness to differences in symptoms in the postoperative period and to collect utilities throughout the disease course. This targeted approach was selected over a comprehensive longitudinal assessment of all possible domains of health-related quality of life because it was more consonant with the clinical questions being asked in the study.

The cost analysis is designed to estimate the *difference* in cost between the two treatment arms rather than to tabulate all costs incurred by study patients. Consequently, data collection is focused on costs associated with the initial surgical therapy and early and late complications of surgery, such as early readmissions or late bowel obstructions due to adhesions. In keeping with the standard Cancer and Leukemia Group B (CALGB) approach to economic analyses alongside clinical trials, this cost analysis is resource based. Data on the number of medical resources consumed (including hospital days, intensive care unit days, operating room time, and surgical/laparoscopic supplies)

are recorded for all trial patients. The difference between study arms in the number of resources consumed will be calculated for each category.

In addition, hospital bills are being collected from three sites that vary in geographic location, size, and teaching status. Billing data will be used to generate estimates of the cost multipliers for each of these resource units. These estimates will be trial specific. For example, the estimate of the mean charge for a hospital day for patients in this trial will reflect not only the hotel component of that stay but also the charges for laboratory tests, medications, radiologic procedures, etc., performed on an average day. Costs will be estimated from charges using hospital- and department-specific ratios of costs to charges. The result of the cost analysis will therefore be an estimate of how many more resources one treatment consumes than the other as well as an estimate of the magnitude of the associated additional cost.

Cost–utility will be determined by dividing this cost difference by the observed difference in quality-adjusted survival between the arms. If the more expensive procedure proves to result in superior quality of life (or less likely, length of life), it will be useful to know whether the extra costs are justified by the extra benefits. Because the quality-of-life component of the trial includes the collection of utility data from patients, estimation of the cost–utility of laparoscopic-assisted colectomy in comparison with open colectomy from trial data will not require any additional data collection beyond that already planned to assess the cost and quality of life in the two trial arms.

Quality-adjusted survival will be calculated from observed survival data and prospectively collected utilites using the method of Q-TWiST (quality-adjusted time without symptoms of disease and toxicity of treatment) (*27,28*). The Q-TWiST method proceeds in four steps as follows: 1) Health states likely to be characterized by different levels of quality of life are identified for the specific disease under study and the treatments being evaluated; 2) overall survival time of patients in the study is partitioned into these health states; 3) the total time spent in each health state by patients in each arm of the trial is multiplied by a utility coefficient or weight reflecting the quality of life reported by patients in that health state; and 4) the average quality-adjusted survival in each trial arm is determined by summing the weighted survival times.

Five health states will be included in the Q-TWiST analysis: 1) the perioperative period (periop), 2) adjuvant chemotherapy (chemo), 3) TWiST (time without symptoms and toxicity), 4) late complications (comp), and 5) relapse (rel). TWiST will be defined as the time from the end of the perioperative period to recurrence or study closure, whichever occurs first, less the duration of adjuvant chemotherapy. Relapse will include time from the diagnosis of recurrence to death or study closure.

Utility weights will be calculated separately for each arm of the trial and will be obtained directly from trial patients using the single-item, 0-100 rating scale of overall quality of life, transformed and recalibrated to a 0-1 scale. Q-TWiST quality-adjusted survival for each treatment group will be calculated by multiplying time spent by trial patients in each health state by the mean patient-reported utility (u) for that state according to the following formula:

$$\text{Q-TWiST} = u_{\text{periop}} \times 30\ \text{d} + u_{\text{chemo}} \times \text{duration}_{\text{chemo}} + u_{\text{TWiST}} \times \text{duration}_{\text{TWiST}} + u_{\text{comp}} \times \text{duration}_{\text{comp}} + u_{\text{rel}} \times \text{duration}_{\text{rel}}.$$

The data will also be presented graphically for each treatment arm as shown in the hypothetical plot in Fig. 2.

Pilot data suggest that laparoscopic-assisted colectomy may be more expensive than open colectomy despite shorter hospital lengths of stay because of increased operative times and costs (*29*). At best, laparoscopic-assisted colectomy may be expected to produce equivalent survival and better quality of life. The cost–utility analysis is designed to permit a determination of whether the magnitude of any observed quality-of-life benefit is sufficient to justify the additional cost of the minimally invasive approach.

Much additional methodologic work is needed to identify optimal approaches to measuring utilities in the clinical trial setting, to refine techniques for calculating quality-adjusted survival from observed survival data, and to establish standards for what constitutes reasonable cost–utility ratios. It is critical that this work proceed as quickly as possible. The relevant data must be available to legislators and regulators if quality-of-life considerations are to receive the recognition they deserve as these individuals make tough choices about how to spend our shrinking health care dollar.

References

(*1*) Detsky AS, Naglie IG. A clinician's guide to cost-effectiveness analysis [*see* comment citation in Medline]. Ann Intern Med 1990;113:147-54.

(*2*) Eisenberg JM. Clinical economics. JAMA 1989;262:2879-86.

(*3*) Karnofsky DA, Abelman WH, Craver LF, Burchenal JH. The use of nitrogen mustards in palliative treatment of carcinoma. Cancer 1948;1:634-56.

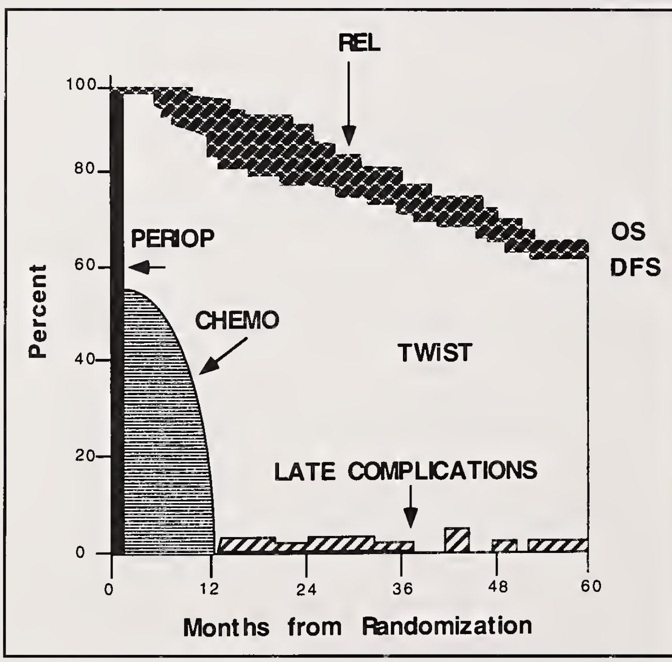

Fig. 2. Hypothetical plot of quality-adjusted survival determined using the Q-TWiST (quality-adjusted time without symptoms of disease and toxicity of treatment) methodology for one arm of the randomized trial of laparoscopic-assisted colectomy versus open colectomy. PERIOP = perioperative, CHEMO = chemotherapy, REL = relapse, OS = overall survival, and DFS = disease-free survival.

(4) von Neumann J, Morgenstern O. Theory of games and economic behavior. New York: Wiley, 1953.
(5) Ganz PA, Schag CA, Lee JJ, Sim MS. The CARES: a generic measure of health-related quality of life for patients with cancer. Qual Life Res 1992;1:19-29.
(6) Schipper H, Clinch J, McMurray A, Levitt M. Measuring the quality of life of cancer patients: The Functional Living Index-Cancer: development and validation. J Clin Oncol 1984;2:472-83.
(7) Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, et al. Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. J Chronic Dis 1981;34:585-97.
(8) Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality-of-life assessment in cancer clinical trials. Recent Results Cancer Res 1988;111:231-49.
(9) Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. Med Care 1988;26:724-35.
(10) Priestman TJ, Baum M. Evaluation of quality of life in patients receiving treatment for advanced breast cancer. Lancet 1976;1:899-900.
(11) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy Scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.
(12) Andrykowski MA, Altmaier EM, Barnett RL, Otis ML, Gingrich R, Henslee-Downey PJ. The quality of life in adult survivors of allogeneic bone marrow transplantation. Correlates and comparison with matched renal transplant recipients. Transplantation 1990; 50:339-406.
(13) Levy SM, Herberman RB, Lee JK, Lippman ME, d'Angelo T. Breast conservation versus mastectomy: distress sequelae as function of choice [*see* comment citation in Medline]. J Clin Oncol 1989;7:367-75.
(14) Weinstein MC, Stason WB. Foundations of cost-effectiveness for health and medical practices. N Engl J Med 1977;296:716-21.
(15) Tsevat J, Weeks JC, Guadagnoli E, Tosteson AN, Mangione CM, Pliskin JS, et al. Using health-related quality-of-life information: clinical encounters, clinical trials, and health policy. J Gen Intern Med 1994;9:576-82.
(16) Torrance GW. Measurement of health state utilities for economic appraisal. A review. J Health Econ 1985;5:1-30.
(17) Sumner W, Nease R, Littenberg B. U-titer: a utility assessment tool. In: Proceedings of Symposium on Computer Applications in Medical Care. Washington, DC, 1992:701-5.
(18) Morss SE, Lenert LA, Faustman WO. The side effects of antipsychotic drugs and patients' quality of life: patient education and preference assessment with computers and multimedia. In: Proceedings of Symposium on Computer Applications in Medical Care. Washington, DC, 1994:17-21.
(19) Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. Socioecon Planning Sci 1976;10:129-36.
(20) O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of time-tradeoff utilities and rating scale values in cancer patients and their relatives: evidence for a possible plateau relationship. J Med Decis Making 1995;15:132-7.
(21) Kaplan RM, Anderson JP. A general health policy model: update and applications. Health Serv Res 1988;23:203-35.
(22) Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R. Multi-attribute preference functions for a comprehensive health status classification system. Paper 92-18. Hamilton, ON, Canada: Centre for Health Economics and Policy Analysis, McMaster University, 1992.
(23) Euroqol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.
(24) Weeks J, O'Leary J, Fairclough D, Paltiel D, Weinstein M. The "Q-tility Index": a new tool for assessing health-related quality of life and utilities in clinical trials and clinical practice. Proc ASCO 1994;13:436.
(25) Nelson H, Weeks JC, Wieand HS. Proposed phase III trial comparing laparoscopic-assisted colectomy versus open colectomy for colon cancer. Monogr Natl Cancer Inst 1995;19:51-6.
(26) McCorkle R, Quint-Benoliel J. Symptom distress, current concerns and mood disturbance after diagnosis of life-threatening disease. Soc Sci Med 1983;17:431-8.
(27) Gelber RD, Goldhirsch A, Cavalli F. Quality-of-life-adjusted evaluation of adjuvant therapies for operable breast cancer. The International Breast Cancer Study Group [*see* comment citation in Medline]. Ann Intern Med 1991;114:621-8.
(28) Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. Stat Med 1990;9:1259-76.
(29) Senagore AJ, Luchtefeld MA, MacKeigan JM, Mazier WP. Open colectomy versus laparoscopic colectomy: are there differences? Am Surg 1993; 59:549-53; discussion 553-4.

Note

Supported in part by the American Society of Clinical Oncology.

Measuring Quality of Life in Culturally Diverse Populations

*Richard B. Warnecke, Carol Estwing Ferrans, Timothy P. Johnson, Gloria Chapa-Resendez, Diane P. O'Rourke, Noel Chávez, Susan Dudas, Eva D. Smith, Lucy Mártinez Schallmoser, Roger P. Hand, Thomas Lad**

If the U.S. National Cancer Institute is to meet its goals for reducing cancer incidence and mortality, it has become increasingly evident that there will have to be a major focus on minority populations. Cancer statistics consistently indicate that, compared with the cancer incidence and survival in the general population, the incidence is higher and the survival lower in minority groups. In recognition of these statistics, it is now mandated that all federally sponsored research explicitly include females and minorities or provide a specific explanation for why they have been excluded. One result of these concerns has been the need to develop valid assessment instruments, appropriate for ethnic minority populations, for use in clinical trials and other research.

Despite earlier assertions to the contrary (*1,2*), researchers increasingly recognize the pitfalls of uncritically applying standard measures in studies of minority populations (*3-14*). Culturally mediated differences in cognition and interpretation are now regarded as responsible for many of the systematic differences that have been observed in cross-cultural surveys (*7,13*). In a series of studies specifically relevant to quality of life, for example, it has been well documented that there are significant cross-cultural differences in how quality of life is assessed (*15-19*), in the perception and reporting of pain (*20-23*), and in illness behavior (*24,25*). Moreover, other ongoing research, funded by the National Center for Health Statistics, has recently shown the effects of culture and educational attainment on a whole range of standard health questions taken from the Health Interview Survey and other major federal health surveys (*26*).

The present article describes the initial phase of a two-phase project designed to assess the applicability of the Ferrans and Powers Quality of Life Index (QLI) (*27*) among cancer patients with a high school education or less who were selected from two minority populations. In the research described here, cognitive methods were used to 1) identify the meaningfulness of specific items in four content domains of the QLI for both male and female adult African-American and Mexican-American cancer patients with low educational levels and 2) determine the capability of cancer patients to form judgments about their satisfaction with and the importance they attribute to life aspects associated with quality of life. Based on these assessments, this article describes a process by which the QLI was modified to be more appropriate for these patients. In a second phase of this project, the resulting instrument is being tested in clinical trials with a larger sample of patients with the same ethnic and educational attributes.

Culture and Assessment of Quality of Life

Virtually every cooperative group now incorporates quality-of-life end points into their clinical trial research protocols. This approach represents a significant departure from the traditional approach in which the end points have been tumor response and duration of survival, while the patient's well-being was not considered (*28,29*).

The increasing pressure to ensure that ethnic minorities are included in clinical trials and the growing interest in adding quality of life as an outcome require more explicit attention to how cultural, ethnic, religious, and other values influence judgments about quality of life (*30-33*). In addition, there is growing agreement that the patient's perceptions provide the most important indicator of quality of life (*24-26*). The Ferrans and Powers QLI (*27,34-36*) was designed to take into account individual values in measuring quality of life. Quality of life in the QLI is defined as "a person's sense of well-being that stems from satisfaction or dissatisfaction with areas of life that are important to him/her" (*27*). Judgments about satisfaction were selected because satisfaction implies a cognitive judgment of experiences based on comparisons of desired and actual conditions of life (*28,30,37*).

Conceptual Basis of the Ferrans and Powers QLI

Conceptually, for the Ferrans and Powers QLI, quality of life is multidimensional, composed of the following four domains:

**Affiliations of authors:* R. B. Warnecke, T. P. Johnson, G. Chapa-Resendez, D. P. O'Rourke (Survey Research Laboratory, College of Urban Planning and Public Affairs), C. E. Ferrans, S. Dudas, E. D. Smith (College of Nursing), N. Chávez (Community Health Sciences, School of Public Health), L. M. Schallmoser, T. Lad (Loyola University of Chicago), R. P. Hand (College of Medicine), University of Illinois at Chicago.

Correspondence to: Richard B. Warnecke, Ph.D., Survey Research Laboratory, College of Urban Planning and Public Affairs, University of Illinois at Chicago, P.O. Box 6905 M/C 336, Chicago, IL 60680.

See "Notes" section following "References."

1) health and functioning, 2) social and economic, 3) psychological/spiritual, and 4) family (*36*). Thirty-three specific life aspects comprise these four domains (Table 1). The QLI was derived from an extensive literature review, measurement based on patient interviews, and then factor analysis. Extensive psychometric assessment of the QLI indicates strong validity and reliability across both patients and diseases.

Category Fallacy and Quality-of-Life Measurement

Despite the strength of psychometric properties, the patient populations used to develop the QLI were primarily well-educated middle and upper middle class individuals. While these populations did include African-American and Hispanic patients, these patients tended to be fairly well educated and, in the case of Hispanics, acculturated to the point of being bilingual. Although a Spanish-language version of the QLI was evaluated with at least one group of patients, the kind of testing done here was not done in this or other earlier versions (*38*).

Measures developed by and for middle and upper middle class respondents are increasingly believed to misrepresent the thoughts, feelings, and behaviors of individuals from segments of the population where poverty is common, where reading ability is limited, and, in the case of Latinos, where ability to speak English is limited or nonexistent (*39-41*). This misrepresentation due to variation in cultural understanding of the question has been described as the "category fallacy." It is a problem with much of the health-related survey data collected in the United States among these populations (*26*).

The category fallacy results from the failure to distinguish between *etic* concepts that are truly universal and accepted across multiple cultural groups and *emic* concepts that have meaning only within a specific cultural group or socioeconomic context. When an *emic* construct is used as if it were *etic*, the resulting construct is described as *pseudoetic*, and the measure results in a category fallacy (*42*). Whether a concept is *etic* or *emic* is believed to be related to its level of abstraction (*43,44*).

Triandis and Marín (*45*) proposed a strategy that emphasizes distinguishing between culturally specific measures (*emic*) and those that are universally relevant (*etic*). They called for using probes designed to assess and understand when unique aspects of the culture influence interpretation and response to questions and when the concept underlying the question is culturally transcendent (*45*). The "*emic* + *etic*" methodology avoids the pitfalls of the pseudoetic approaches that adversely affect questionnaire design. This is the approach used in this study.

Methods

The Cognitive Strategy

Cognitive research on the validity of responses to survey questions has identified four steps in the response process (*46-48*). Although not every step is followed by every respondent when answering every question, the four steps are 1) question interpretation, 2) information retrieval, 3) judgment formation, and 4) response editing.

Question interpretation. Culturally influenced language and interpretation differences are likely to influence how respondents understand questions dealing with quality of life (*49,50*). In some instances, the language into which a question is translated does not contain the concept. In other instances, cultural mediation or cultural experience influences the meaning or validity of the question. In either case, when the respondent replies, the reply is not to the same question that has been asked; hence, it is not valid (*51,52*). Cognitive assessment seeks to understand what the question means to the respondent and, if the meaning is different from that intended by the questioner, to guide the choice of more culturally or educationally appropriate wording.

Information retrieval. By this process, either an answer or information relevant for constructing an answer is retrieved from memory. One major area of inquiry has been how question wording can be modified to provide cues to facilitate recall (*53*). Individuals differ in how they access their memories for such information (*54*). Recall of information about regular, recurring events is likely to be "semantic," involving schemas in which information about classes of events is retained in memory rather than information about specific events. In contrast, events that occur sporadically or very occasionally are subject to "episodic" recall where recalled information is focused on specific episodes or events (*54-56*). It has been observed that semantic schema are likely to be culturally influenced by the individual's community or larger culture. Accuracy of recall may also be culturally related (*57-59*).

Judgment formation. Based on information retrieved from memory, judgment formation is an important aspect of attitude formation. It is the process by which the importance of events and satisfaction with one's current life status are translated into an assessment of quality of life (*27,28,30,34-37*). Most often when a respondent is asked about the value attributed to a life aspect such as an event, experience, or action, the response is a synthesis of information retrieved from memory about relevant experience. The more frequently such information

Table 1. Specific aspects of quality-of-life domains

Specific aspects by domain			
Health and functioning domain	Social and economic domain	Psychological/ spiritual domain	Family domain
Usefulness to others	Standard of living	Life satisfaction	Family happiness
Physical independence	Financial independence	Happiness	Children
Responsibilities	Home	Self	Spouse
Own health	Job/unemployment	Goal achievement	Family health
Stress and worry	Neighborhood	Peace of mind	
Leisure activities	Friends	Personal appearance	
Retirement	Emotional support	Faith in God	
Travel	Education	Control over life	
Live a long life			
Sex life			
Health care			
Pain			
Energy (fatigue)			

is used, the more accessible it is in memory; hence, the more readily it is used for developing judgments (*60,61*).

Perhaps of most relevance to quality-of-life research are the findings suggesting that, when individuals search their memory for specific information, they use contextual cues such as references to specific persons, events, or locations. These cues lead the respondent to access particular sets of memories that are likely to mediate the response (*62,63*). These retrieval cues can only help access memories that have been previously encoded (*64*). Because of the overwhelming amount of information that is available, not all of it may be important enough to be encoded in memory (*51*).

Questions about satisfaction and importance assume that the individual has stored in memory relevant experiences that will be available for forming judgments about the importance of and current satisfaction with the life aspects that are the point of each question. If there are no memories of specific and relevant events to cue the patient's responses, however, the actual responses to questions about how much the patient is satisfied with a particular life aspect or how important it is may be based on motivation to be a "good respondent." Hence, the response may be subject to editing rather than reflecting the respondent's true assessment. (*See* the following section on response editing.) Cultural conditioning may also directly influence judgment formation. For example, some research on responses to health opinion surveys suggests that Hispanics in the United States are more fatalistic than Anglo respondents regarding cancer (*65*) because of the culturally specific concept of *fatalismo*, or the belief that little can be done to alter one's fate. Other research suggests that cultural variation in the likelihood of probabilistic thinking (i.e., the ability to express thoughts in terms of uncertainty) may also influence how judgments are formed (*66,67*).

Finally, the validity of scales requires a common frame of reference for mapping judgments onto a common metric. For example, African-American and Hispanic survey respondents are less likely than Anglo-Americans to qualify their answers on rating scales, whereas Asians are less likely to prefer extreme responses (*68-71*). Preference for extreme versus cautious response styles has been interpreted as being a consequence of cultural variation in emphasis on sincerity versus modesty in social interaction (*71,72*). Use of modifiers tends to increase among Hispanics with acculturation (*70*).

Response editing. This editing is a commonly encountered phenomenon when survey respondents feel that certain answers are more socially desirable than others (*73*). For example, socially desirable behaviors such as exercise and nutrition are frequently overreported, whereas such undesirable behaviors as drinking or smoking are frequently underreported. Available information suggests that definitions of socially desirable behavior vary culturally (*51,74-77*). Being Mexican and being a member of a minority group have been correlated with giving socially desirable responses (*78,79*). Socially desirable response patterns are compatible with the commonly observed pattern of social interaction in Hispanic cultures referred to as *simpátia*, or the expectation that interpersonal relations will be guided by harmony and the absence of confrontation (*80*). Such cultural expectations also seem to influence Asian survey respondents (*81*).

A related phenomenon is respondent acquiescence, or the tendency to agree with a statement regardless of its content (*82*). Acquiescence is observed as a strategy of self-presentation most commonly, although not universally (*23,24*), among low-status Hispanics and African-Americans (*3,68,70,72,79,82*). Alternatively, it may be that acquiescence occurs because of too much *emic* question content, leading respondents who are unsure about what is being asked to "play it safe" and acquiesce rather than to look foolish or admit they do not understand the question (*26*).

Editing also occurs in situations where there is social or cultural distance between the interviewer and respondent because of ethnicity, gender, educational level, or other status indicators (*83-96*). There is also evidence that bilingual respondents may answer differently, depending on the language used by the questioner and the cultural significance of the question (*76,96*), which may affect the tendency for acquiescence, social desirability, cultural understanding, or cross-cultural accommodation (*76,97*). It may also be that language variation produces differences in response cues that affect recall and/or judgment (*98*).

Cognitive Methods

The Ferrans and Powers QLI was evaluated by use of cognitive probes designed to explore whether African-American and Mexican-American cancer patients varied in the way in which they understood the various components of the QLI, how they retrieved information and formed judgments regarding the importance of and their satisfaction with each life aspect, and whether they edited their responses. Individual respondents are selected because they have educational or cultural characteristics that might affect how they may interpret the questionnaire content. Thus, they are recruited and interviewed.

The cognitive interview has been developed over the last decade in research on questionnaire design by teams of survey methodologists and cognitive psychologists (*99-102*). During the cognitive interview, the respondent is asked the question to be evaluated. Once he or she answers the question, then standard probes or follow-up questions are used to explore understanding, retrieval, judgment formation, and editing effects in the answer. These probes help the respondent reconstruct or "think aloud" about the thought processes used to respond to the question.

The interaction between the cognitive interviewing process and the questionnaire is iterative. As more is learned about how these processes affect response through the cognitive process, the question content is revised. Further interviews are conducted until there is apparent consensus among respondents regarding the meaning of individual questions, which is the final objective of the process. When the final revisions are made, the questionnaire is finalized and pretested. Thus, we kept using a question in the interviews until the responses became redundant. When respondents had interpretive problems, questions were revised (sometimes several times) and then retested until no further linguistic or interpretive problems were identified.

Before we began the cognitive interviews, the questionnaire was reviewed by a reading literacy laboratory in the College of Education at the University of Illinois. The purpose of this process was to revise or eliminate any questions where overall reading level might interfere with the cognitive processes being evaluated through the "think-aloud" probes.

Patient Selection

The purpose of this study was to assess the validity of the Ferrans and Powers QLI among African-American and Mexican-American patients with low education and, in the case of the Mexican-American subjects, poorly acculturated. African-American patients selected for cognitive interviews were recruited from outpatient clinics affiliated with the University of Illinois and Mount Sinai hospitals in Chicago. They were eligible if they had a high school education or less, and the range in education among subjects varied from third grade to high school diploma or equivalent. Interviews were conducted between February and September 1994 with 23 African-American patients (nine females and 14 males).

Fifteen Mexican-American patients (11 females and four males), selected according to the same criteria as used for the African-American subjects, were interviewed in October 1994 at The University of Texas M. D. Anderson Cancer Center in Houston, TX. These interviews were conducted in Spanish and, with patients from whom Spanish was the primary language, by bilingual interviewers.

Patients were selected with the knowledge and consent of their attending physicians. While awaiting chemotherapy, patients were recruited by nursing staff in the clinics. All respondents were informed that their participation in the interview was voluntary. Respondents were given an honorarium for participating.

Results

The analysis focused on the content of each specific element of life quality in the four domains of the Ferrans and Powers QLI (Table 1) and on the overall scaling used to obtain the respondents' ratings of the importance of and satisfaction with each element. We will first consider the domain content and then the scales.

Most of the problematic questions related to education and reading level. In the results reported below, the questions were initially evaluated and altered during interviews with the African-American patients who were interviewed first. This pattern was deliberate because of the costs associated with translation into Spanish. Thus, we attempted to resolve the issues related to literacy and the interaction between cognitive responses and education before we translated the questionnaire.

When we evaluated the questions using cognitive probes with Mexican-American patients at The University of Texas M. D.

Anderson Cancer Center, we discovered additional problems due to linguistic issues. For the most part, however, the educational level needed by the Mexican-American patients to understand and form judgments about the questions was the same as that required by the African-American patients. Question wordings that were changed as a result of the interviews conducted in Spanish were retested in English on African-American patients. In the summary of results below, questions that were problematic for Hispanic patients are identified in the text.

Domain 1: Health and Functioning

Seven of the 13 items related to health and functioning required some revision based on the readability assessment and cognitive interviewing process.

Interpreting the question was the problem most commonly encountered by the respondents. Three questions were revised based on the literacy evaluation. These questions were as follows:

1) **Original:** *How satisfied are you with your usefulness to others?*
Revised: *How satisfied are you with how useful you are to others?*

2) **Original:** *How satisfied are you with your leisure time activities?*
Revised: *How satisfied are you with the things you do for fun?*

3) **Original:** *How satisfied are you with your potential for a happy old age/retirement?*
Revised: *How satisfied are you with your chances for a happy future?*

Problems of understanding the underlying concept emerged from the probing process in a fourth question:

4) **Original:** *How satisfied are you with your physical independence?*
Probe: *What do the words "physical independence" mean to you?*

The responses to the probe indicated that the term "physical independence" was being interpreted as financial independence or as not being reliable for others who depended on the respondent. In one interview, the respondent simply said he or she did not know what the term meant.

Revised: *How satisfied are you with your ability to take care of yourself without help?*

With that revision, subsequent respondents quickly achieved consensus. Respondents were clearly able to describe the "ability to do things without help." The revised form of the question was incorporated into the QLI.

5) **Original:** *How satisfied are you with the amount of stress or worries in your life?*
Probes: *Can you tell me in your own words what this question is asking about?*
What does [the word] "stress" mean to you?
What does [the word] "worries" mean to you?
Did you answer this question in terms of stress, worries, or both?
Would it have been easier to answer, harder to answer, or about the same if we did not include both worry and stress in the same question?

To the African-American respondents, there was considerable overlap in the meaning of the terms "worries" and "stress," but the results from the probes for this question indicated that using "worries" produced greater validity than when the term "stress" was used. Our decision regarding validity was based on the number of respondents who indicated that they understood what the question was asking during the probes and who based their responses on that understanding. Moreover, during the Spanish language interviews, it became clear that, linguistically, there is no term in Mexican Spanish for "stress" and using synonyms for stress changed the translation of the question.

Revised: *How satisfied are you with the amount of worries in your life?*

Information retrieval problems occurred with one question from this domain.

6) **Original:** *How satisfied are you with your ability to travel on vacations?*
Probes: *What determines your ability to travel?*
Do you take vacations?
If you wanted to take a vacation and had the money to do so, would your health be good enough to do so?
What do you think we mean by vacation?

The concepts of vacation and travel for pleasure had no equivalent meanings that would allow retranslation or reformulation of the question that asked about these things. Neither the concept of "vacation" nor the concept of "travel for pleasure" had meaning for the African-American and Mexican-American respondents because neither had relevance to their lifestyle or experience. For example, if they traveled, it was to visit family and only for a family emergency. The question was dropped when it became clear that there was no way the concept could be written that would cue relevant memories on which to base a response. The relevant elements of the concept were covered by the question discussed above: "How satisfied are you with the things you do for fun?"

Judgment formation issues also arose in one question where the probing indicated variation in the anchoring point associated with the age of the respondent.

7) **Original:** *How satisfied are you with your potential to live a long time?*
Probes: *What do you think we mean by "your potential to live a long time"?*
What do you consider "a long time" to be?

The problem of establishing a common scale for forming judgments arose because the response to the original question depended on the age of the respondent. In point of fact, this issue is probably relevant to all who use this scale, regardless of education or acculturation. In response to the probes, older respondents replied that they already had lived a long time; younger respondents responded

in terms of the future. We revised the wording to cue respondents to respond in terms of whatever expectations about continued longevity they might have.

Revised: *How satisfied are you with your chance of living to the age you would like?*

Domain 2: Social and Economic

Four of the 10 questions in this domain were revised following probing.

Question interpretation was a problem with two items in this domain. One item was revised following assessment for reading level as follows:

1) **Original:** *How satisfied are you with your financial independence?*
Revised: *How satisfied are you with how well you can take care of your financial needs?*

For the second revised question, the problem was clearly interpretation, and the question created problems for both the African-American and Mexican-American patients.

2) **Original:** *How satisfied are you with your standard of living?*
Probes: *What do you think we mean by "standard of living"?*
What kinds of things do you think about in answering this question about your standard of living?

In response to the probes, several African-American respondents interpreted the question as addressing a moral issue, "standards of living." Moreover, there was no straightforward translation into Spanish of the concept standard of living. The item was dropped because the elements that it was intended to address were adequately covered by other questions dealing with financial needs and satisfaction with home and neighborhood.

Information retrieval was combined with question interpretation in the two remaining problematic questions from this domain.

3) **Original:** *How satisfied are you with the amount of emotional support you get from others?*
Probe: *What do you think we mean by "emotional support"?*

In response to the probes, especially the query about "others," it was clear that, as written, the question did not offer specific cues about whose support was relevant; some thought "others" referred to friends, and some thought the term referred to family. Two questions were ultimately used: one about family and one about friends. This procedure produced consensus to the probes. Thus, the final question wording used was as follows:

Revised: *How satisfied are you with the emotional support you get from your family?*
How satisfied are you with the emotional support you get from people other than your family?

4) **Original:** *How satisfied are you with your home?*
Probes: *In your own words, what do you think this question is asking about your home?*
Why are you [satisfied/dissatisfied] with your home?
What things about your home did you think about when answering this question?

The last probe, requiring information retrieval, indicated problems with this item. The question was designed to address the physical aspects of the home environment. As the respondents "thought aloud" about the things about their homes that influenced their judgments regarding importance and satisfaction, they described the ambience, especially interactions with children in the home and the neighborhood. Finally, at least one respondent did not interpret the question as applicable to apartment dwellers.

Two strategies improved consensus around this question. First, the question was relocated in the QLI to follow other questions that asked specifically about satisfaction with children and the neighborhood. By placing these items before the home question, the respondents were cued that we wanted them to think about *other* aspects of their home *besides* children and the neighborhood. Second, we rewrote the question to refer to several types of dwelling.

Revised: *How satisfied are you with your home, apartment, or place where you live?*

Domain 3: Psychological/Spiritual

There were no problems with this domain among the African-American patients. An item, regarding "personal belief in God," proved quite wordy in the Spanish version. When it was retranslated without the word "personal," it worked well in both English and Spanish.

1) **Original:** *How satisfied are you with your personal faith in God?*
Revised: *How satisfied are you with your faith in God? (Su fe en Dios?)*

Domain 4: Family

Question interpretation caused problems with one question in this domain. It was first evident in the Spanish translation. On review, however, it was also a problem with the English version.

1) **Original:** *How satisfied are you with your relationship with your spouse/significant other?*
Probes: *What parts of the relationship did you think about when you answered this question?*
Has your relationship changed in any way since you got cancer?

In response to the probes, both the African-American and Mexican-American respondents indicated that they thought about the sexual aspects of the relationship. The Spanish translation cued Mexican-American respondents to think about the sexual aspects of the relationship because the term "relationship" in Spanish is translated as "relaciónes," which specifically means sexual intercourse. As the QLI was originally designed,

the questions about satisfaction and importance of sex life followed the questions about satisfaction with one's spouse and the importance of that relationship. The ordering of these two questions was changed so that the sexual satisfaction question preceded the spousal satisfaction question. This order cued respondents that the question on spousal satisfaction did not refer to the sexual relationship. There was some confusion in both languages about who was a "significant other," so the wording was changed.

Reworded: *How satisfied are you with your spouse, lover, or partner?*

Measurement Scales

Respondents were asked to form judgments about their satisfaction with each of the 33 life aspect items (*see* Table 1) and then to weight the importance of each item. Both of these scaling tasks required the respondents to form judgments and format their responses using bipolar scales ranging from "very satisfied" to "very dissatisfied" and from "very important" to "very unimportant," respectively.

Early in the cognitive interview process, it became apparent that the African-American respondents experienced difficulty with the importance and satisfaction scales and the task of recording their judgments using the labeled, six-item, bipolar scales. The labels were presumed to form an equal interval scale that discriminated between levels of satisfaction and importance. Upon further examination, the intervals were not perceived as discrete but rather as overlapping.

The satisfaction scale asked: "How satisfied are you with . . . [each life aspect]?" It then asked the respondent to describe their satisfaction by selecting one of the following terms: very satisfied, moderately satisfied, slightly satisfied, slightly dissatisfied, moderately dissatisfied, and very dissatisfied.

The importance scale asked: "How important is . . . [each life aspect]?"

Respondents were again asked to select from a series of phrases: very important, moderately important, slightly important, slightly unimportant, moderately unimportant, and very unimportant.

The respondents did not understand the bipolar scaling that they were being asked to perform. To assess the nature of the problem, a thermometer scaled from "0" to "100" was presented to 23 additional African-American and Spanish-speaking respondents who were selected on the basis of educational level but who were not cancer patients. The respondents were then asked to locate a variety of descriptors of "satisfaction" and "importance" on the thermometer (Fig. 1).

Inasmuch as the process and results were the same for both scales, we will report the results of the "importance" scale here. On the thermometer presented to respondents, "0" was labeled "as unimportant as something could ever be," "100" was labeled "as important as something could ever be," and "50" was labeled "neither important nor unimportant." Respondents were given a series of terms that they were asked to place on the thermometer between 0 and 100. The following terms reflected importance: "very important," "totally important," "important," "somewhat important," "moderately important," "a little important," "fairly important," "slightly important," "neither important nor unimportant," "not very important," "somewhat unimportant," "fairly unimportant," "moderately unimportant," "slightly unimportant," "a little unimportant," "totally unimportant," "unimportant," "not at all important," and "very unimportant."

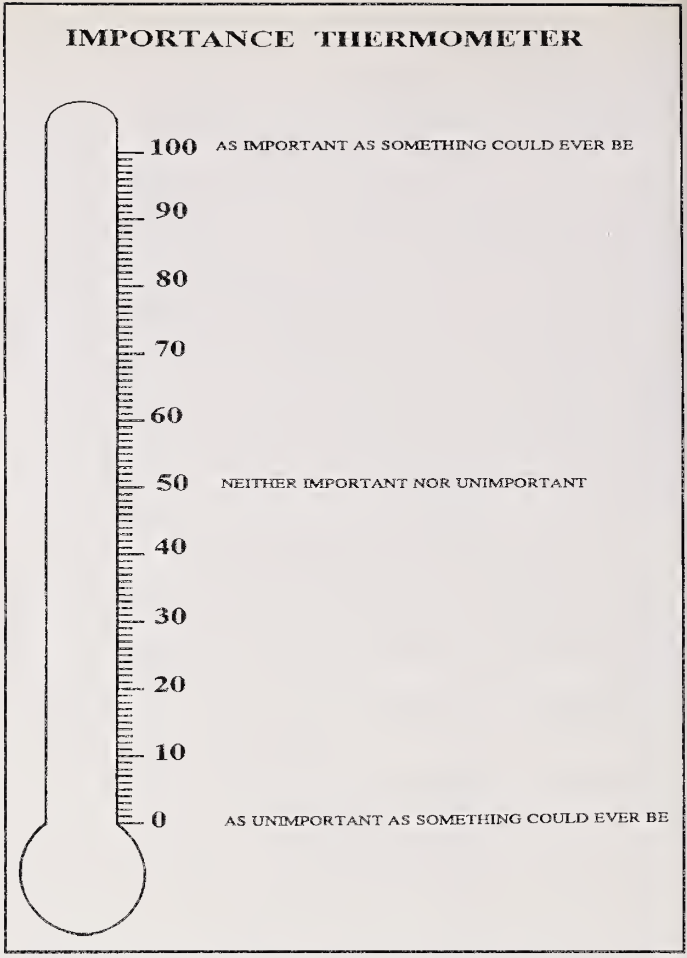

Fig. 1. Thermometer used by respondents to describe "satisfaction" or "importance."

Table 2 presents the range of values assigned to the key terms actually used in the scale. Two things are evident from this table. First, based on the range of values, there was considerable overlap in the numeric rating of each descriptor. That is, the range for "a little important" (90-05) overlapped with the range

Table 2. Range of responses to scale of importance or unimportance

Scale	Range of responses
Scale of importance	
Very important	100-90
Important	100-80
Somewhat important	100-45
A little important	90-05
Scale of unimportance	
Very unimportant	0-70
Unimportant	0-50
Somewhat unimportant	10-95
A little unimportant	0-60

assigned to "important" (100-80). "A little important" (90-05) almost completely encompassed "somewhat important" (100-45). The same pattern could be observed for the various characterizations of "unimportant." The range of values assigned to "very unimportant" (0-70) totally encompassed the ranges for "unimportant" (0-50) and for "a little unimportant" (0-60), which in turn had a wider range of values than "unimportant."

The second point is that the respondents did not see "very important" and "very unimportant" as polar opposites on a single scale from 0 to 100. On the contrary, the respondents treated the range of importance and the range of unimportance as two separate and overlapping scales. For example, the range of values assigned to "very important" was 100-90. However, the range assigned to "very unimportant" (0-70) overlapped "a little important" (90-05). "Somewhat unimportant" (10-95) overlapped "very important" (100-90), "important" (100-80), "somewhat important" (100-45), and "a little important" (90-05). If the point of this scale is to discriminate, the bipolar scale clearly did not achieve that objective with these subjects.

An additional matter that arose among the Spanish-speaking respondents was that there are no direct Spanish linguistic equivalents for either "unimportant" or "dissatisfied." Hence, the negative pole of each scale as previously translated was not cognitively understood in the way it was worded in English. Bilingual respondents solved this problem by devising a translation from Spanish into English that was consistent with the scale. Thus, although *sin importancia*, for example, is directly translated from Spanish as "not important" or "without importance," the bilingual respondents understood that to be equivalent to unimportant. Because there was no direct translation for "unimportant" which is the negative pole of the scale, however, the Spanish speakers did not understand the bipolar contrast. Thus, the only solution was to change the English to match the Spanish and to use a single dimensional scale anchored by "very important" and "not at all important" and to use a similar 6-point scale anchored by "very satisfied" and "not at all satisfied." The graphic used for the self-administered version of the QLI and as an aid in the interview version is a bar graph indicating an "amount" of satisfaction or importance from 1 to 6 (Fig. 2). When the scale was retested using the graphics, subjects with low education were able to complete the entire interview without problems; moreover, they used the entire range of the scale, including the midpoint.[1]

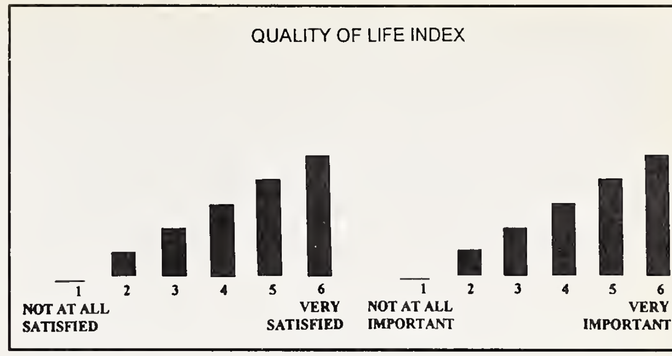

Fig. 2. Bar graph indicating an "amount" of "satisfaction" or "importance" from 1 to 6.

Discussion and Conclusions

Several important points are evident from this analysis. First, there was clear evidence of problems during three of the four cognitive stages. (No editing problems were encountered.) The greatest problems were with question interpretation. In part, these problems resulted from questions that were written to require a level of verbal comprehension that was too high for the likely respondents. These problems were easily corrected by rewriting the question to require lower levels of verbal comprehension. Other problems of question interpretation came from language. In some instances, the English concept was not meaningful linguistically in Spanish. Of importance was the fact that, when these questions were asked of bilingual respondents, as they had been in earlier evaluations of the QLI (*38*), the respondents were able to translate them into English, arrive at an answer, and back-translate their responses into Spanish. If the respondent did not speak any English, these questions either were not understood and were left unanswered or, more often, were answered as a way of acquiescing to the interviewer. This was the only form of editing that we were able to observe in the way these scales were used.

Some concepts, such as standard of living and traveling on vacation, had meaning but no recallable memory content. That is, the respondents could not retrieve information relevant to making an answer because they had never experienced the events in question. These questions had to be dropped.

In some cases, the meaning was not well specified in the item as written. By paying close attention to the relationship between questions and to the wording, the meaning could be more clearly specified. This was the case with the question regarding relationship with one's spouse. This question was initially interpreted as a question about satisfaction with the sexual aspects of the relationship until we changed the order of the question to follow the specific question regarding satisfaction with sex life. Once the order was changed, it was clear that the two questions referred to different aspects of the spousal relationship.

Finally, there was the response scale. It is clear that bipolar and labeled scales were less successful with these respondents than unipolar scales where only the end points are labeled. Using a graphic also aided response. With it, the scaling task was understandable even to respondents with very low levels of verbal skills, and these respondents were able to use the entire scale including the midpoints appropriately. This result illustrates the importance of clarifying ambiguous or vague quantifiers such as "moderately satisfied" for respondents with weak verbal skills.

Another important finding was the consistent evidence that respondents will answer questions that are *emic* in order not to reveal their inability to understand a question. That is, they will "satisfice." It is important, therefore, at a minimum to translate and back-translate, both of which were done with the original scale. When scales are used with respondents who are not bilingual, however, it is also important to test the scales with respondents who are not bilingual. Only when the scale was tested with persons who spoke only Spanish did some of the linguistic problems emerge. Overall, the results demonstrate the impor-

tance of this kind of assessment in all scales before they are translated to a different language or administered to a population different from the one for whom the instrument was developed.

Finally, our experience with both the individual items and the scales indicated the importance of both conceptual equivalence across cultural groups and the way the question is phrased. The failure to consider the conceptual equivalence and wording across language groups will limit the generalizability and validity of the results. It was most efficient to deal with the potential problems in order, starting with reading level, then conceptual understandability, and finally linguistic problems. It is also important to recheck changed questions with the groups where consensus as to meaning has been established. Thus, in the final phase, we interviewed our final group of African-American respondents *after* we did the interviews in Houston to ensure that the changed items were still valid in the African-American population.

It is also important to emphasize that our data represent data on samples of patients who were selected because they represented the ethnic groups on which the QLI was to be tested and because they had relatively low educational levels. Their responses cannot be generalized to any larger population. The selection of these two patient groups was driven by the funding source that specifically requested proposals to examine the generalizability of quality-of-life scales to these populations.

The QLI is now being evaluated on larger populations of patients at the University of Illinois Hospitals and Clinics and at The University of Texas M. D. Anderson Cancer Center. It is also being administered at several points during the therapy, so that we will have data on how the scale changes during treatment. Following this trial, we intend to conduct psychometric studies on the data to establish norms for these populations. These data will be compared with data obtained from a variety of other populations of patients on whom the scale has already been evaluated. In all probability, the QLI will be revised.

References

(*1*) Freeman DM. A note on interviewing Mexican Americans. Soc Sci Q 1969;49:909-18.

(*2*) Welch S, Comer J, Steinman M. Interviewing in a Mexican American community: an investigation of some potential sources of response bias. Public Opin Q 1973;37:115-26.

(*3*) Aday LA, Chiu GY, Anderson R. Methodological issues in health care surveys of the Spanish heritage population. Am J Public Health 1980;70:367-74.

(*4*) Flaherty JA. Appropriate and inappropriate research methodologies for Hispanic mental health. In: Gaviria M, Arana JD, editors. Health and behavior: research agenda for Hispanics. The Simón Bolivar Research Monogr Series No. 1. Chicago: Univ Illinois at Chicago, 1987:177-86.

(*5*) Marsella AJ. Thoughts on cross-cultural studies on the epidemiology of depression. Cult Med Psychiatry 1978;2:343-57.

(*6*) Manfredi C, Lacey L, Warnecke R, Balch G, Allen K. Complementary information from survey data and focus group insights. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Lancaster, PA, 1990.

(*7*) Marín G, Marín BV. Research with Hispanic Populations. Newbury Park (CA): Sage Publications, 1989.

(*8*) Milburn NG, Gary LE, Booth JA, Brown DR. Conducting epidemiological research in a minority community: methodological considerations. J Commun Psychol 1991;19:3-12.

(*9*) Montero D. Research among racial and cultural minorities: an overview. J Social Issues 1977;33:1-10.

(*10*) Myers V. Survey methods for minority populations. J Social Issues 1977;33:11-9.

(*11*) Salber EJ, Beza AG. The health interview survey and minority health. Med Care 1980;28:319-26.

(*12*) Vernon SW, Roberts RE, Lee ES. Ethnic status and participation in longitudinal health surveys. Am J Epidemiol 1984;119:99-113.

(*13*) Word CO. Cross-cultural methods for survey research in black areas. J Black Psychol 1977;3:72-87.

(*14*) Zmud JP, Arce CH. Language cue management: techniques for improving response rates and data quality in surveys of Hispanics. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ, 1991.

(*15*) Rogler IH. The meaning of culturally sensitive resesarch in mental health. Am J Psychiatry 1989;146:296-303.

(*16*) Vaughan DA, Kashner JB, Stork WA, Richards M. A structural model of subjective well-being. Social Indicators Res 1985;16:315-32.

(*17*) Mukherjee M, Ray A, Rajyalaksme C. Physical Quality of Life Index. Social Indicators Res 1979;6:283-92.

(*18*) Mukherjee R. On the quality of life in India. Social Indicators Res 1981;9:455-76.

(*19*) Mastekaasa A, Moum T. The perceived quality of life in Norway. Social Indicators Res 1984;14:385-420.

(*20*) Morse JM, Morse RM. Cultural variation in the inference of pain. J Cross-Cultural Psychol 1988;19:232-42.

(*21*) Zborowski M. Cultural components in responses to pain. J Social Issues 1952;8:16-30.

(*22*) Zborowski M. People in pain. San Francisco: Jossey-Bass, 1969.

(*23*) Zola IK. Culture and symptoms—an analysis of patients' presenting complaints. Am Sociol Rev 1966;31:615-30.

(*24*) Good B, Kleinman A. Culture and anxiety. In: Tuma AH, Maser JD, editors. Anxiety and anxiety disorders. Hillsdale (NJ): Lawrence Erlbaum, 1985.

(*25*) Harwood A, editor. Ethnicity and medical care. Cambridge (MA): Harvard Univ Press, 1981.

(*26*) Johnson T, O'Rourke D, Chavez N, Sudman S, Warnecke R, Lacey L, et al. Social cognition and responses to survey questions among culturally diverse populations. Presented at the International Conference on Survey Measurement and Process Quality, Bristol, UK, 1995.

(*27*) Ferrans CE. Development of a quality of life index for patients with cancer. Oncol Nurs Forum 1990;17(3 Suppl):15-9.

(*28*) Caires N, List M, Lansky S. Article reviewed: Quality of life: what is it? How should it be measured? Oncology 1988;2:76.

(*29*) Tannock IF, Boyer M. When is a cancer treatment worthwhile? N Engl J Med 1990;323:989-90.

(*30*) Andrews F, Withey S. Social indicators of well-being. New York: Plenum Press, 1976.

(*31*) Calman K. Definitions and dimensions of quality of life. In: Aronson N, Beckmann O, editors. The quality of life of cancer patients. New York: Raven Press, 1987:1-9.

(*32*) Ferrans C. Quality of life of breast cancer survivors. Oncol Nurs Forum 1990;(216 Suppl).

(*33*) Oleson M. Subjectively perceived quality of life. Image J Nurs Sch 1990;22:187-90.

(*34*) Ferrans CE, Powers MJ. Quality of life index: development and psychometric properties. ANS Adv Nurs Sci 1985;8:15-24.

(*35*) Ferrans CE, Powers MJ. Employment potential of hemodialysis patients. Nurs Res 1985;34:273-7.

(*36*) Ferrans CE, Powers MJ. Psychometric assessment of the Quality of Life Index. Res Nurs Health 1992;15:29-38.

(*37*) Campbell A, Converse C, Rodgers W. The quality of American life. New York: Russell Sage, 1976.

(*38*) Martinez-Schallmoser L. Perinatal depressive symptoms, quality of life, social support, and risk factors in Mexican-American women [dissertation]. Chicago: Univ Illinois, 1992.

(*39*) Azibo D. Understanding the proper and improper usage of the comparative research framework. J Black Psychol 1988;15:81-91.

(*40*) Adams-Esquivel H. Conceptual adaptation vs. back-translation of multilingual instruments: how to increase the accuracy and actionability of multilingual surveys. Proceedings of the American Association of Public Opinion Research, May 1991. Ann Arbor (MI): Am Assoc Public Opin Res, 1991.

(*41*) Berry JW. On cross-cultural comparability. Int J Psychol 1969;4:207-29.

(*42*) Triandis HC. The analysis of subjective culture. New York: Wiley-Interscience, 1972.

(*43*) Scheuch EK. The cross-cultural use of sample surveys: problems of comparability. Historical Soc Res 1993;18:104-38.

(*44*) Triandis HC. Some universals of social behavior. Personality and Soc Psychol 1978;4:1-16.

(*45*) Triandis HC, Marín G. Etic plus emic versus pseudoetic: a test of a basic assumption of contemporary cross-cultural psychology. J Cross-Cultural Psychol 1983;14:489-500.

(46) Strack F, Martin LL. Thinking, judging, and communicating: a process account of context effects in attitude surveys. In: Hippler HJ, Schwarz N, Sudman S, editors. Social information processing and survey methodology. New York: Springer-Verlag, 1987:123-48.
(47) Tourangeau R. Attitude measurement: A cognitive perspective. In: Hippler HJ, Schwarz N, Sudman S, editors. Social information processing and survey methodology. New York: Springer-Verlag, 1987:149-62.
(48) Tourangeau R, Rasinski KA. Cognitive processes underlying context effects in attitude measurement. Psychol Bull 1988;103:209-314.
(49) Angel R, Gronfein W. The use of subjective information on statistical models. Am Sociol Rev 1988;53:464-73.
(50) Dressler WW, Viteri FE, Chavez A, Grell AC, DosSantos JE. Comparative research in social epidemiology: measurement issues. Ethn Dis 1991;1:379-93.
(51) Angel R, Thoits P. The impact of culture on the cognitive structure of illness. Cult Med Psychiatry 1987;11:465-94.
(52) Pepitone A, Triandis HC. On the universality of social psychological theory. J Cross-Cultural Psychol 1988;18:471-97.
(53) Sudman S, Schwarz N. Autobiographical memory and the validity of retrospective reports. New York: Springer-Verlag, 1994.
(54) Tulving E. Elements of episodic memory. New York: Oxford Univ Press, 1983.
(55) Blair E, Burton S. Cognitive processes used by survey respondents to answer behavioral frequency questions. J Consumer Res 1987;14:280-8.
(56) Menon G. Judgments of behavioral frequencies: memory search and retrieval strategies. In: Schwarz N, Sudman S, editors. Autobiographical memory and the validity of retrospective reports. New York: Springer-Verlag, 1994:161-72.
(57) Engle PL, Lumpkin JB. How accurate are time-use reports? Effects of cognitive enhancement and cultural differences on recall accuracy. Applied Cognitive Psychol 1992;6:141-59.
(58) Fendrich M, Vaughn CM. Diminished lifetime substance use over time. Pub Opin Q 1994;58:96-123.
(59) Warnecke RB, Havlicek PL, Manfredi C. Awareness and use of screening by older-aged persons. In: Yancik R, Carbone PP, Patterson WB, Skel K, Terry WD, editors. Perspectives on prevention and treatment of cancer in the elderly. New York: Raven Press, 1983:275-87.
(60) Dovidio JF, Fazio RH. New technologies for the direct and indirect assessment of attitudes. In: Tanur J, editor. Questions about questions: inquiries into the cognitive bases of surveys. New York: Russell Sage Foundation, 1992:204-37.
(61) Sudman S, Bradburn N, Schwarz N. Thinking about answers: the applications of cognitive processes to survey methodology. San Francisco: Jossey-Bass, 1995.
(62) Brewer WF. Memory for randomly sampled autobiographical events. In: Neisser U, Winograd E, editors. Remembering reconsidered: ecological and traditional approaches to the study of memory. Cambridge: Cambridge Univ Press, 1988:21-90.
(63) Wagenaar WA. My memory: a study of autobiographical memory over six years. Cognitive Psychol 1985:18:225-52.
(64) Strube G. Answering survey questions: the role of memory. In: Hippler HJ, Schwarz N, Sudman S, editors. Social information processing and survey methodology. New York: Springer-Verlag, 1987:86-101.
(65) Pérez-Stable, EJ, Sabogal F, Otero-Sabogal R, Hiatt RA, McPhee SJ. Misconceptions about cancer among Latinos and Anglos. JAMA 1992;268:3219-23.
(66) Vaughn E, Nordenstam B. The perception of environmental risks among ethnically diverse groups. J Cross-Cultural Psychol 1991;22:29-60.
(67) Wright GN, Phillips LD, Whalley PC, Choo GT, Ng KO, Tan I, et al. Cultural differences in probabilistic thinking. J Cross-Cultural Psychol 1978;9:285-99.
(68) Bachman JG, O'Malley PM. Black-white differences in self esteem: are they affected by response styles? Am J Sociol 1984;90:624-39.
(69) Hui CH, Triandis HC. Effects of culture and response format on extreme response style. J Cross-Cultural Psychol 1989;20:269-309.
(70) Marín G, Gamba RJ, Marín BV. Extreme response style and acquiescence among Hispanics: the role of acculturation and education. J Cross-Cultural Psychol 1992;23:498-509.
(71) Zax M, Takahashi S. Cultural influences on response style: comparisons of Japanese and American college students. J Soc Psychol 1967;71:3-10.
(72) Bachman JG, O'Malley PM. Yea-saying, nay-saying, and going to extremes: black-white differences in response styles. Pub Opin Q 1984;48:491-509.
(73) Bradburn NM, Sudman S. Improving interview method and questionnaire design; response effects to threatening questions in survey research. San Francisco: Jossey-Bass, 1979.
(74) Dohrenwend B. Social status and psychological disorder: an issue of substance and an issue of method. Am Sociol Rev 1966;31:14-34.
(75) Groves RM. A total survey error approach to AIDS-related survey research. In: Fowler FJ, editor. Conference Proceedings on Health Survey Research Methods. Washington, DC: DHHS Publ No. (PHS) 89-3447, 1989:265-70.
(76) Marín G, Triandis HC, Betancourt H, Kasima Y. Ethnic affirmation versus social desirability: explaining discrepancies in bilingual responses to a questionnaire. J Cross-Cultural Psychol 1983;14:173-86.
(77) Presser S. Pretesting: A neglected aspect of survey research. In: Fowler FJ, editor. Conference Proceedings on Health Survey Research Methods. Washington, DC: DHHS Publ No. (PHS) 89-3447, 1989:35-37.
(78) Ross CE, Mirowsky J. The worst place and the best face. Social Forces 1983;62:529-36.
(79) Ross CE, Mirowsky J. Socially-desirable response and acquiescence in a cross-cultural survey of mental health. J Health Soc Behav 1984;25:189-97.
(80) Triandis HC, Marín G, Lisansky J, Betancourt H. *Simpátia* as a cultural script for Hispanics. J Pers Soc Psychol 1984; 47:1363-75.
(81) Deutcher I. Asking questions: linguistic comparability. In: Warwick D, Osherson S, editors. Comparative research methods. Englewood Cliffs (NJ): Prentice Hall, 1973:163-86.
(82) Schuman H, Presser S. Questions and answers in attitude surveys: experiments on question form, wording and context. San Diego: Academic Press, 1981.
(83) Carr LG. The scale items and acquiescence. Am Sociol Rev 1971;36:287-93.
(84) Alers JO. Interviewer effects on survey response in an Andean estate. Int J Comparative Sociol 1970;11:208-19.
(85) Anderson BA, Silver BD, Abramson PR. The effects of the race of the interviewer on race-related attitudes of black respondents in the SRC/CPS National Election Studies. Pub Opin Q 1988;52:289-324.
(86) Andersen RM, Mullner RM, Cornelius LJ. Black-white differences in health status: methods or substance? Milbank Q 1987;65(Suppl 1):72-99.
(87) Campbell BA. Race-of-interviewer effects among southern adolescents. Pub Opin Q 1981;45:231-44.
(88) Cotter PR, Cohen J, Coulter PB. Race-of-interviewer effects in telephone interviews. Pub Opin Q 1982;46:278-86.
(89) Finkel SE, Guterbock TM, Borg MJ. Race-of-interviewer effects in a preelection poll: Virginia 1989. Pub Opin Q 1991;55:313-30.
(90) Hatchett S, Schuman H. White respondents and race-of-interviewer effects. Pub Opin Q 1975;39:523-28.
(91) Reese SD, Danielson WA, Shoemaker PJ, Chang TK, Hsu HL. Ethnicity-of-interviewer effects among Mexican-Americans and Anglos. Pub Opin Q 1986;50:503-72.
(92) Schaeffer NC. Evaluation race-of-interviewer effects in a national survey. Sociol Methods Res 1980;18:400-19.
(93) Schuman H, Converse J. The effects of black and white interviewers on black responses in 1968. Pub Opin Q 1971;35:44-68.
(94) Sudman S, Bradburn NM. Response effects in surveys: a review and synthesis. Chicago: Aldine, 1974.
(95) Summers G, Hammonds A. Effect of racial characteristics of investigator on self-enumerated responses to a Negro prejudice scale. Soc Forces 1966;48:525-8.
(96) Weeks MF, Moore RP. Ethnicity of interviewer effects on ethnic respondents. Pub Opin Q 1981;45:245-9.
(97) Bond MH, Yang KS. Ethnic affirmation versus cross-cultural accommodation: the variable impact of questionnaire language on Chinese bilingual in Hong Kong. J Cross-Cultural Psychol 1982;13:169-85.
(98) Feldman RH. The effect of administrator and language on traditional-modern attitudes among Gusii students in Kenya. J Soc Psychol 1975;96:141-2.
(99) Jabine TB, Staf ML, Tanur JM, Tourangeau R. Cognitive aspects of survey methodology: building a bridge between disciplines. Washington, DC: Natl Acad Press, 1984.
(100) Jobe JB, Mingay DB. Cognition and survey measurement: history and overview. Applied Cognitive Psychol 1991;5:175-92.
(101) Lessler J, Tourangeau R, Salter W. Questionnaire design in the Cognitive Research Laboratory. National Center for Health Statistics, Vital and Health Statistics. DHHS Publ No. (PHS) 89-1501, 1989.
(102) Royston PN. Using intensive interviews to evaluate questions. In: Fowler FJ, editor. Conference Proceedings on Health Survey Research Methods. Washington, DC: DHHS Publ No. (PHS) 89-3447, 1989:3-7.

Notes

[1]We thank David Cella, Ph.D., Rush University, Chicago, IL, for suggesting this approach.

Supported by Public Health Service grant R01CA61698-02 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services (C. E. Ferrans, Principal Investigator).

We thank the following individuals for their help: Arthur Boddie, Henry Briele, Tapas Das Gupta, Thomas Lad, Michael Regan, Roobollah Sharifi, Michael Warso, Barry Wenig, and the staff at the participating clinics (University of Illinois Hospital and Clinics, Cook County Hospital and Clinics, West Side Veterans Administration Hospital and Clinics College of Medicine) and Robert Levin and his staff (Mount Sinai Hospital and Clinics). We also thank the following research assistants who helped with both respondent recruitment and the cognitive interviews: Shaunda Bonds, Charles Bright, Ana N. Chapa, Francisco Perez, and Jonathan VanGeest.

Empirically Selected Instruments for Measuring Quality-of-Life Dimensions in Culturally Diverse Populations

*Frank Baker, David Jodrey, James Zabora, Charlene Douglas, Patricia Fernandez-Kelly**

We describe a process for developing and testing the cultural equivalence of quality-of-life (QOL) instruments that may be used across culturally diverse populations. QOL instruments dealing with satisfaction with various life domains, psychological distress, and physical health and functioning were reviewed by African-American and Hispanic community advisory boards, translated into Spanish and back-translated to ensure translation adequacy, administered to samples of 100 patients from each of the ethnic minority populations by indigenous nurse interviewers, and examined for psychometric adequacy. Ten QOL measures showed adequate reliability and validity for further use in the assessment of QOL with African-American and Hispanic patients. Three other measures failed to meet the defined standards. A dimension shown to be particularly difficult to address across culturally diverse groups is family functioning. Procedures for achieving cultural equivalence of QOL measures have been shown to be practical and productive. Measures are identified that may be used with some confidence to assess varied dimensions of QOL with culturally diverse groups. [Monogr Natl Cancer Inst 1996; 20:39-47]

As the number of individuals surviving cancer and other life-threatening diseases has increased during the last decade, there has been increasing recognition of the importance of conducting research on the psychosocial adaptation and quality of life (QOL) of long-term survivors (*1,2*). There has also been increased acceptance that QOL should be considered a major outcome criterion for the assessment of the medical effectiveness of demanding treatments.

The extent to which interest in QOL in cancer patients has increased dramatically is demonstrated by the large number of reviews of this area [e.g., (*3-11*)]. These reviews generally agree that QOL is an important criterion in studies of the consequences of treatment of cancer patients. They also agree that QOL is a multidimensional concept that includes psychological, functional, and social dimensions and that its assessment should include self-report measures from patients. However, the reviews also point out that QOL measurement still poses a variety of problems that trouble those responsible for assessing the effectiveness of cancer treatments. Among these problems is the question of appropriateness of these measures for use with patients from diverse cultural and socioeconomic backgrounds.

Significant evidence exists that minority populations experience higher mortality rates and suffer higher incidences of diseases than other populations in the United States (*12,13*). Given differential rates in incidence and mortality, one can conclude that minority populations may lack adequate access to the health care system. Among the various factors that interfere with receiving adequate health care are the generally recognized factors related to adequacy of employment, income, and health insurance coverage. Cultural factors, including attitudes, beliefs, customs, and practices, also affect whether these population groups seek care and how they participate in and respond to care. Barriers also exist related to difficulties with the majority language and educational requirements.

These problems not only contribute to underservice of patients from minority and low socioeconomic groups, but also affect their inclusion in clinical trials in cancer treatment and supportive care. QOL measurement has grown in acceptance as an important component in medical decision making and in the evaluation of medical treatment effectiveness. The absence of normative data from special populations for QOL measures has limited the use of these measures in clinical trials.

Problems in Multicultural Research

The problem of developing tests and measures that can be used across culturally diverse groups is not limited to medical research; much relevant experience exists in psychological and anthropological research over a considerable period of time. In the United States, the concern for developing adequate tests and measures to use cross-culturally was greatly stimulated by social and political developments in the second half of this century. However, the general problem of developing cross-cultural

**Affiliations of authors:* F. Baker, D. Jodrey, Department of Environmental Health Sciences, The Johns Hopkins University School of Hygiene and Public Health, Baltimore, MD; J. Zabora, The Johns Hopkins Oncology Center and The Johns Hopkins School of Medicine, Baltimore; C. Douglas, Center for Health Policy, George Mason University School of Nursing and Health Science, Fairfax, VA; P. Fernandez-Kelly, The Johns Hopkins University Institute for Policy Studies and the Department of Sociology, The Johns Hopkins University.

Correspondence to present address: Frank Baker, Ph.D., American Cancer Society, 1599 Clifton Rd., N.E., Atlanta, GA 30329-4251.

See "Notes" section following "References."

psychological tests was recognized as early as 1910 during the attempt to develop tests for use in research on the comparative abilities of the large groups of immigrants coming to this country at the turn of the century (*14*).

Those who have criticized multicultural research have identified a number of problems with regard to the instruments used. They include the following: 1) Instruments have been developed on the white middle-class population (*15*), and 2) the conceptual base used in creating these instruments has been Eurocentric, and it may be inappropriate to use these instruments with nonwhite, non-middle-class respondents (*16-19*).

Achieving Cultural Equivalence

It is possible to reduce cultural distortion in order to make comparisons, such as evaluating the effects of treatment on cancer patients in culturally diverse groups. This requires a process of adapting instruments to achieve what Flaherty et al. (*20*) have called "cultural equivalence."

Flaherty et al. (*20*) have identified the following five major dimensions as relevant to establishing the cultural equivalence of measures to be used cross-culturally: 1) *content equivalence*—whether the content of each item is relevant to the phenomena of each culture being studied; 2) *semantic equivalence*—whether the meaning of each item is the same in each culture after translation into the language and idiom (written or oral) of each culture; 3) *technical equivalence*—whether the method of assessment (e.g., pencil and paper, interview) is comparable in each culture with respect to the data it yields; 4) *criteria equivalence*—whether the interpretation of the measurement of the variable remains the same when compared with the norm for each culture studied; and 5) *conceptual equivalence*—whether the instrument is measuring the same theoretical construct in each culture. These five types of equivalence that are necessary for achieving measures that work across cultural boundaries offer a basis for examining QOL instrument development for use with culturally diverse populations.

Culturally Equivalent QOL Instruments

Funded by the National Cancer Institute, we have been developing instruments for assessing QOL that will provide comparable data in cancer patients from culturally diverse populations, including African-Americans and Hispanic Americans who vary in levels of literacy and socioeconomic status. During the first phase of this 3-year project, existing QOL measures have been examined and modified as necessary to achieve cultural equivalence.

The research design has been structured to deal with each of the five issues of cultural equivalence identified by Flaherty et al. (*20*).

To achieve content equivalence, advisory groups from the two racial/ethnic minority groups reviewed the measures and rated their content.

To establish semantic equivalence, the English language versions of measures were translated into Spanish by one bilingual person, and the Spanish version was back-translated into English by others. To deal with differences in colloquial language used in different national groups of Spanish speakers, an attempt was made to use "broadcast Spanish," the type of Spanish used on the radio and television. These translations and back-translations were reviewed by an Hispanic American advisory board, including members with Central American, South American, and Puerto Rican backgrounds.

With regard to technical equivalence, these two advisory groups of individuals familiar with the subcultures of relevance provided input on the procedures of administration of the measures and critiqued the response formats.

To reduce the effect of interviewer differences on responses to the QOL questions, the suggestions of Choi and Comstock (*21*) were used and included selecting interviewers with similar characteristics and backgrounds, training them adequately and conducting periodic field assessment of their performance, simplifying the questions and reducing the number of possible responses per question, and allocating various types of subjects to the interviewers as uniformly as possible. The measures were administered as part of a face-to-face interview with large community samples of patients from Baltimore, MD, Washington, DC, and Northern Virginia by nurses from the same racial/ethnic minority groups, who also completed rating forms on how well the measures were understood and whether there were any problems in acceptability, content, format, or wording.

Criteria equivalence and conceptual equivalence are being examined through psychometric and other statistical analyses of the data obtained from samples from the two culturally different groups in relation to data obtained from earlier administrations of the instruments with other groups.

Three-Phase Study

Our overall research design involves a three-phase process. The first phase, as outlined above, involves review of available instruments, initial testing with community samples of African-American and Hispanic patients with chronic disease, and psychometric evaluation of the performance of these measures. The second phase involves administration of the measures that performed adequately in the first phase to cancer patients from the special populations. The third phase will consist of testing the measures resulting from the first two phases in clinical trials with special population cancer patients. Data from the first phase are currently available as a basis for comparing instruments for measuring QOL in culturally diverse populations.

In the first-phase study, the initial step was to assess the adequacy of these measures by having them reviewed by advisory boards made up of people knowledgeable about the culture and language of the special population groups. Instruments in this review process included measures of satisfaction with various life areas, global QOL, psychological well-being, depression, mood states, level of physical functioning, health status, pain, family functioning, and current concerns. The selection of instruments was based on a review of the literature and the researchers' experience in conducting QOL research with cancer patients and other groups of patients with chronic disease in varied community settings during the past 15 years.

QOL measures were reviewed by each of the two advisory boards as described above and then administered to a sample of

each population. A sample of 100 African-American and a sample of 100 Hispanic patients with various chronic diseases who resided in Baltimore, Washington, or Northern Virginia were recruited to complete the 10 QOL measures that had survived the advisory board review process.

The QOL scales were administered by specially trained interviewers recruited from the African-American and Hispanic patient population groups. The interviewers were predominantly nurses who were familiar with the respective communities. The Hispanic interviewers were all fluent in Spanish. Interviewers were given a training manual that described each of the instruments and instructed the interviewers with regard to the administration of these scales and the accompanying interview questions. Separate group-training sessions were held for the African-American and Hispanic interviewers. The booklet that presented the interview questions and the structured scales for the Hispanic patients provided both English (on the left) and Spanish (on the right) versions of the questions on facing pages. Hispanic patients were allowed to respond in English if they preferred to, but the great majority chose to use Spanish in the interview and in answering the scales.

Patients were recruited through hospitals, public health clinics, churches, and a variety of other community organizations. The subjects gave informed consent before the interviews. After completing the interview/questionnaire, they were paid a nominal sum for their time and transportation costs. Interviews were conducted at the patients' homes or at several offices that were made available at clinics, churches, and other community organizations in Baltimore, Washington, and Northern Virginia.

Study Samples

Table 1 presents the demographic characteristics of the African-American and Hispanic samples studied. The initial sample of 100 African-American patients included approximately equal proportions of males and females, with an overall mean age of 54.27 years (range, 24-84 years). Almost two fifths (38%) of the patients included in this sample were married or living with someone. With regard to educational level, 3% had only an elementary education, 6% had only a middle school education, 32% had some high school, 27% had graduated from high school, 21% had some college or university education, 7% were college graduates, and 4% did not provide this information.

With regard to the initial sample of 100 Hispanic patients, 49% were male. The mean age of Hispanic patients was somewhat lower (47.39 years). Almost two thirds of this group were married or living with someone. With regard to citizenship status, 17% were U.S. citizens, 54% were permanent residents (i.e., had green cards), 25% were without documents or in some other "informal or temporary" status, and 4% did not provide such information. With regard to education, the Hispanic sample reported less schooling than the African-American sample.

QOL Instruments

Ten QOL measures have been successfully put through the advisory board review and translation process and administered to the initial samples of 100 African-American and 100 Hispanic patients. Table 2 presents basic data on these measures, including the following characteristics: 1) number of items, 2) type of response format, 3) availability of alternative versions, 4) rating of ease of response, 5) acceptability rating, and 6) reliability (Cronbach alpha). The measures that have been tested for use with culturally diverse populations are described below. Scales 1 and 2 are general measures of QOL; scales 3-6 assess dimensions of psychological distress; scales 7-10 deal with physical health and functioning dimensions of QOL.

1) The *Satisfaction With Life Domains Scale for Cancer (SLDS-C)* is a broad measure of QOL that asks about multiple aspects of life. It is based on an earlier Satisfaction With Life Domains Scale developed by Baker et al. (*22,23*) to assess the QOL of chronic psychiatric patients. The SLDS-C asks those completing the scale to indicate their satisfaction with a number of different life domains relevant to the QOL of cancer patients using a picture response format. Respondents are asked to express their feelings about 17 life areas by choosing one of seven faces, ranging from a "delighted" face with a large smile (scored 7) to a "very unhappy" face with a deep, down-turned frown (scored 1), a response format shown by Andrews and Withey (*24*) in national QOL surveys to be easily used and well accepted by most respondents. The "smiley face" response format, which may be presented to the respondent on a card without printed words, has been shown to work well with interviewees who have limited language or conceptual capabilities. The development of the earlier Satisfaction With Life Domains Scale for Mental Illness (SLDS-MI) was undertaken to obtain a QOL measure that could be used in the evaluation of community-based support services for deinstitutionalized mental patients, who were low in levels of socioeconomic status,

Table 1. Demographic characteristics of the samples

	Sample	
Characteristic	African-American, % (n = 100)	Hispanic, % (n = 100)
Sex		
Male	48	49
Female	52	51
Age, y		
Mean	54.27	47.39
Range	24-84	18-82
Marital status		
Married/living with someone	38	62
Widowed	23	6
Separated/divorced	20	12
Single	19	20
Highest educational level attained		
Elementary	3	21
Middle school	6	7
Some high school	32	28
High school graduate	27	18
Some college	21	15
College graduate	7	9
Missing*	4	2

*Did not provide information.

Table 2. QOL measures for culturally diverse populations

Instrument	No. of items	Response format	Alternative versions	Ease of response	Acceptability	Reliability,* African-American/ Hispanic
General QOL						
Satisfaction With Life Domains Scale for Cancer	17	Smiley faces	General bone marrow transplant Breast Mentally ill	+++	+++	.93/.90
Cantril Ladder of Life	1	Ladder	Past Present Future	+++	+++	NA
Psychologic distress						
Center for Epidemiologic Studies—Depression Scale	20	No. of days in past week felt this way		+++	+++	.89/.91
Shacham Profile of Mood States	37	How much felt like adjective 0-4	Total negative mood	++	++	.95/.96
			Subscales			
			Tension			.81/.83
			Depression			.91/.92
			Anger			.87/.97
			Fatigue			.85/.92
			Vigor			.75/.86
			Confusion			.74/.81
Bradburn Positive Affect Scale	5	No Sometimes Often	Affect balance scale	+	+	.65/.73
Bradburn Negative Affect Scale	5	No Sometimes Often	Affect balance scale	+	+	.80/.75
Physical health and functioning						
Self-rated Karnofsky Performance Scale	7	Check one statement		++	+	
MOS† SF-20 Physical Functioning Scale	6			+	+	
MOS† SF-20 General Health Perceptions Scale	5			++	++	
MOS† SF-20 Bodily Pain Scale	1			+++	+++	

*The reliability data presented here are alphas based on administrations of the scales to community samples of 100 African-American and 100 Hispanic patients (Spanish language version used). NA = not applicable.

†Medical Outcomes Study.

literacy, and ability to handle abstractions. The SLDS-C shows only partial overlap in the life domains rated in completing the scale. Sample items include "your relations with friends," "your body," "how comfortable you feel," and "your ability to attain sexual satisfaction."

Evidence of the reliability of the SLDS-C was first obtained from analysis of the responses of a sample of 109 cancer patients. The coefficient alpha for this group was .93, and evidence of the SLDS-C's concurrent validity was also demonstrated on the basis of its correlation with several other QOL measures (*25*). Evidence of its sensitivity to change was obtained by comparing the responses of 64 cancer patients who completed the measure as part of a resurvey 2 years after completing an initial mailed questionnaire. A statistically significant gain in average SLDS-C was observed in the subset of cancer patients who indicated that their health had improved (*25*). The construct validity of the bone marrow transplant version of the scale (which has an additional item that asks about bone marrow transplantation) has been supported by a study showing that the ability of 135 cancer survivors who had had bone marrow transplants to maintain their valued social roles was significantly related to a higher QOL as measured by the 18-item SLDS-BMT version of the scale (*26*).

2) The *Cantril Ladder of Life* is another graphic method for studying QOL. It asks about overall life satisfaction (*27*). Respondents are shown a "ladder of life" with 10 rungs; 0 represents the worst possible life (as the person conceives it), and 10 represents the best possible life. Respondents indicate where they are on the ladder at the present time. Other versions ask patients to indicate where they were in the past (e.g., before they had cancer) and where they expect to be in the future. This method has been described as "self-anchoring," since the respondents establish where they are at a particular time on the ladder of life and can use that judgment as a basis for compatibility rating their lives at other time points. This simple measure has been widely used in the study of life satisfaction.

3) The *Center for Epidemiologic Studies—Depression Scale (CES-D)* is a self-report measure of the frequency of depression rated for the past week (*28*). The CES-D consists of 20 items for which patients are asked to circle a number on a scale of 1-4; 1 is defined as "rarely or none of the time (less than 1 day)," and 4 is defined as "most or all of the time (5-7 days)." The CES-D

was developed initially for use in epidemiologic surveys with the general population, and its use for screening people for symptomatology related to depression has been well established (*29,30*). The CES-D is unusual among the QOL measures discussed here, in that it has already been translated into Spanish by its developers and has undergone reliability and validity testing with general samples of several U.S. ethnic minority groups (*31*). Roberts (*32*) in a comparison of samples of white subjects of non-Hispanic origin, African-Americans, and Mexican-Americans found no differences among the three groups with regard to missing data or alpha coefficient reliability. The CES-D has frequently been used to evaluate depression symptoms in cancer patient populations (*33-35*) and has the advantage over other measures of depression of being less biased by the inclusion of items asking about physical concerns that might be expected to reflect symptoms of cancer or its treatment rather than depression (*36*).

4) The *Profile of Mood States (POMS)* assesses transient, distinct mood states by self-report on an adjective checklist (*37*), yielding an overall score of total negative mood and six factor scores including the following: 1) tension–anxiety, 2) depression–dejection, 3) anger–hostility, 4) fatigue–inertia, 5) vigor–activity, and 6) confusion–bewilderment. The original 65-item POMS has shown internal consistencies ranging from .84 to .95 and test–retest reliabilities ranging from .65 to .74 for 20 days and .43 to .53 for 9 weeks (*37,38*). The version we are using is the shortened 37-item POMS developed with cancer patients by Shacham (*39*), and it has shown correlations with the original full-length scale of .95, indicating stability of the shorter version. The POMS has frequently been used to assess the psychological status of patients with cancer [e.g., (*40-44*)] and has also been employed to examine the psychosocial impact of cancer on the family (*45-48*). Graydon (*49*) found that, for cancer patients (while adjusting for diagnosis and age), the tension–anxiety component was the best predictor of functioning after therapy. Cassileth et al. (*44*) have provided comparative POMS scores for cancer patients and next of kin. In prior research by members of this research group with cancer patients surviving bone marrow transplantation, the obtained alpha reliability coefficient for total negative mood on the short Shacham form of the POMS was .94 (*26*).

5 and 6) The *Bradburn Positive and Negative Affect Scales* comprise a set of 10 questions (five negative and five positive) that ask respondents about their recent affective experiences. Intended by Bradburn (*50*) to be a single measure of psychological well-being, the two five-item clusters have been found to be independent in a number of studies (*23,24,51*) and are often used as separate measures of affect. The two measures, the Positive Affect Scale and the Negative Affect Scale, have been shown to be useful outcome measures for chronically ill patients (*22*).

In previous research by members of this research group with cancer patients surviving bone marrow transplantation, the obtained alpha reliability coefficients were .83 for the Positive Affect Scale and .60 for the Negative Affect Scale (*26*). Other researchers have also found this measure to be a useful one in studying the affect dimension of QOL among cancer patients [e.g., (*52*)].

7) The *Self-Rated Karnofsky Performance Scale (SR-KPS)* is a measure of physical functioning for cancer patients. It was developed to provide a self-report version of the classic physician-rated Karnofsky scale (*53*). Patients rate themselves by a 10-point increment from 40 (low-level functioning requiring help) to 100 (high-level functioning requiring no help). The categories of 10, 20, and 30 have been deleted because patients at these lower levels of functioning would be unable to participate in such a study. In a survey of 70 cancer patients after bone marrow transplantation, the SR-KPS was validated against a physician's ratings using the traditional Karnofsky scale, and statistically significant kappas were obtained (*54*).

8) The *MOS SF-20 Physical Functioning Scale* is from the Medical Outcomes Study (MOS) and is a 20-item short form (*54*), which was designed at the RAND Corporation (Santa Monica, CA) as a quick (<5 minutes) self-administered questionnaire for use in large-scale patient surveys. The alpha obtained for the Physical Functioning Scale in the original study was .86 (*54*). It has been shown to be a useful measure of physical functioning in the study of health and functional status of adult cancer survivors after bone marrow transplantation by Wingard et al. (*53*). Responses range from 1 ("all of the time") to 6 ("none of the time") regarding how much of the time during the last month the respondents' health has limited them in each of six types of activities they can do, ranging from vigorous activities such as "lifting heavy objects, running or participating in strenuous sports," to activities of daily living such as "eating, dressing, bathing or using the toilet."

9) The *MOS SF-20 General Health Perceptions Scale* (*54*) is also from the MOS SF-20 and consists of five overall ratings of current health in general. Four of the items ask the respondents to circle a number from 1 ("definitely true") to 5 ("definitely false") indicating whether each statement is true or false for them. Items include such statements as: "I am somewhat ill" and "My health is excellent." A fifth item asks the respondent to circle a number from 1 ("excellent") to 5 ("poor") in rating how their overall health is at the present time. The alpha obtained for this measure in the original study was .88 (*54*). This scale was successfully employed in the study by Wingard et al. (*53*) in assessing the perceived health status of bone marrow transplant survivors. Although a longer form of the MOS measures (i.e., MOS SF-36) is now available, these scales from the earlier short form have the advantage of simplicity and of having already been used with cancer patients in several studies (*53*).

10) The *MOS SF-20 Bodily Pain Scale* is a single item from the MOS SF-20 (*54*) that asks the following question: How much *bodily* pain have you had during the past week? Respondents are asked to circle a number from 1 ("none") to 5 ("severe"). We used a modified wording of this item that asked how much pain the respondent was "feeling now." This version has been shown to be a simple but useful measure of self-reported level of pain in the study of bone marrow transplant survivors noted above (*53*).

Analysis Plan

Our analysis plan was first to examine the psychometric performance of the various scales. A standard approach to

reliability of multi-item scales is to calculate each scale's coefficient alpha (*55*), which has been described as a measure of internal consistency. Reliability as estimated by the alpha coefficient has been considered to be acceptable for a scale at as low a level as .50 for making group comparisons (*56*), although Nunnally (*57*) has recommended using a standard of reliability of .70 or higher.

Preliminary tests of validity were also conducted with this pilot dataset. Convergent validity was examined in terms of the correlation among measures hypothesized as measuring the same dimension of QOL. Discriminant validity was gauged by the extent to which absolute values of correlations were lower for measures presumed to be assessing different dimensions of QOL than for those concerned with the same dimension. Scales were also assessed in terms of the amount of missing data and inconsistency in factor structure for those scales that were developed as having a particular factor structure.

Results

Rejected Instruments

Some measures that we put through the advisory board review and initial testing with special population samples did not survive the process. Three of these measures are of particular interest.

Our experience with assessing QOL had reminded us that cancer is not just a disease that affects the patient, but, in a sense, it is a "family disease" that has an impact on those who are close to the patient and are affected by his or her ordeal (*58*). After reviewing 11 scales related to family functioning, we selected the 20-item FACES III (Family, Adaptation, and Cohesion Evaluation Scales) developed by Olson et al. (*59*), based on Olson's circumplex model of marital and family systems. Unfortunately, although this instrument was based on surveys of thousands of adults (*60*), it was not well accepted by either the African-American or Hispanic samples of patients whom we asked to complete it. Both the African-American and the Hispanic patient samples particularly objected to items that seemed to assume that children share power with their parents in family decision making, such as the following items: "Parents and children discussed punishment together"; "In solving problems, the children's suggestions were followed"; and "Children had a say in their discipline." Even though the respondents had the option of indicating that the particular behavior almost never happened, many of the subjects rejected these items as inappropriate. As a result, they either refused to answer the scale altogether or left a number of the items unanswered.

The Inventory of Current Concerns (ICC), a classic self-report measure built to assess multiple QOL dimensions (Weisman AE, Worden JW, Sobel HJ: unpublished report), was examined. The ICC is a 72-item, self-report inventory that focuses on the patient's "current concerns." These concerns are categorized into seven areas: 1) health, 2) family, 3) work–finance, 4) friends, 5) religion, 6) existential, and 7) self-appraisal. The ICC got through the review by the two advisory boards, although there were concerns about its length and the wording of some of the items; however, it encountered major difficulties at the testing stage of the study. Missing data were higher for items in the scale than any other scales except the FACES III.

Finally, a third scale that was reviewed for possible use was rejected early in the process of measurement review. In order to ascertain the social desirability response bias, the tendency to give answers that make the respondent look good, a shortened version of the Marlowe–Crowne Social Desirability Scale (M-C SDS) (*61*) was considered for inclusion in our study. However, this 10-item scale was eliminated during the advisory board process because members of the African-American Advisory Board objected to this measure in the strongest terms, suggesting that it was extremely insensitive and implied that minority patients were not trusted to tell the truth. The Hispanic Advisory Board did not object in such strong terms, but they too felt that the measure was somewhat offensive.

Acceptable Measures

Table 2 summarizes the results from the first phase of our research in developing culturally equivalent QOL measures. For each of the 10 instruments, the table presents 1) the number of items in each scale, 2) the nature of each scale's response format, 3) whether there have been alternative versions of the measure developed and/or whether it has subscales, 4) how easy using a particular scale's format was for the special populations we studied, 5) how acceptable each instrument appeared to be, and 6) the coefficient alpha reliability obtained from analysis of the responses to each scale of the initial community samples of 100 African-American and 100 Hispanic patients.

Most of the scales are multi-item scales, except for the Cantril Ladder of Life and the MOS SF-20 Bodily Pain Scale. Four of the scales have alternate forms, and one has specific subscales as well as a total scale score. The SLDS has several alternative forms, but only the general cancer version, the SLDS-C, was examined in this study. The Cantril Ladder of Life has been used to rate one's life as it was in the past and as one expects it to be in the future, but only the versions asking the respondents to rate their lives right now were used. The POMS originally was 65 items long, but we are using the shorter Shacham 37-item form developed with cancer patients; however, both the whole scale score for total negative mood and the six subscale scores were examined here. The Bradburn Positive and Negative Affect Scales were originally added together algebraically to create an Affect Balance Scale score (*50*), but we have treated them as two different scales in our analysis, consistent with the way they have generally been used in studies with cancer patients (*26*).

While all 10 measures were found to be reasonably acceptable to the groups to whom we administered them, the measures scoring the highest average rating of ease of use were the SLDS, the Cantril Ladder of Life, the CES-D, and the MOS SF-20 Bodily Pain Scale, reflecting their one-item simplicity or their use of picture response formats. The interviewers observed that these measures seemed most acceptable to respondents as well. The coefficient alphas ranged from .65 to .93, indicating acceptable reliability for all the measures. The scales with highest coefficient alphas were the SLDS, the CES-D, and the Total POMS. The Bradburn Positive Affect Scale had the lowest alpha levels with both ethnic minority groups.

Table 3. Correlations across QOL measures for African-American patients

	Cantril Ladder of Life (now)	Center for Epidemiologic Studies—Depression Scale	Shacham Profile of Mood States—total	Bradburn Positive Affect Scale	Bradburn Negative Affect Scale	Self-rated Karnofsky Performance Scale	Medical Outcomes Study SF-20 Physical Functioning Scale	Medical Outcomes Study SF-20 General Health Perceptions Scale	Medical Outcomes Study SF-20 Bodily Pain Scale
Satisfaction With Life Domains Scale for Cancer	.59*	–.74*	–.76*	.23†	–.68*	.61*	.55*	.61*	–.48*
Cantril Ladder of Life (now)		–.56*	–.55*	.43*	–.54*	.49*	.44*	.48*	–.39*
Center for Epidemiologic Studies—Depression Scale			.84*	–.30*	.69*	–.48*	–.40*	–.53*	.36*
Shacham Profile of Mood States—total				–.32*	.76*	–.54*	–.47*	–.54*	.36*
Bradburn Positive Affect Scale					–.21†	.28*	.14	.25†	–.20†
Bradburn Negative Affect Scale						–.54*	–.46*	–.54*	.39*
Self-rated Karnofsky Performance Scale							.74*	.65*	–.55*
Medical Outcomes Study SF-20 Physical Functioning Scale								.71*	–.53*
Medical Outcomes Study SF-20 General Health Perceptions Scale									–.49*

*Significance, $P \leq .01$.
†Significance, $P \leq .05$.

Validity of QOL Measures

Tables 3 and 4 present intercorrelation matrices for the African-American and Hispanic samples, respectively. The SLDS shows high correlations (about .5 or above) with every measure considered here, except that, for the African-American sample, the correlation with positive affect is smaller ($r = -.23$). Presumably, this reflects the fact that the SLDS is a comprehensive measure that covers a broad range of QOL domains. Originally, the ICC was to be used to test conversant validity with SLDS-C, but this could not be done because of the ICC's poor performance and low acceptability, as discussed below. Future studies are planned to examine the convergent validity of this multidomain measure with another QOL measure that measures multiple QOL dimensions.

The one-item Cantril Ladder of Life for the present has its highest correlation with the SLDS for the African-American sample and scored highest for the Hispanic sample. Like the other general QOL measure, the SLDS, it shows high negative correlations with the measures of psychological distress.

Establishing convergent and discriminant validity for the measures of psychological distress used in these samples re-

Table 4. Correlations across QOL measures for Hispanic patients

	Cantril Ladder of Life (now)	Center for Epidemiologic Studies—Depression Scale	Shacham Profile of Mood States—total	Bradburn Positive Affect Scale	Bradburn Negative Affect Scale	Self-rated Karnofsky Performance Scale	Medical Outcomes Study SF-20 Physical Functioning Scale	Medical Outcomes Study SF-20 General Health Perceptions Scale	Medical Outcomes Study SF-20 Bodily Pain Scale
Satisfaction With Life Domains Scale for Cancer	.66*	–.72*	–.75*	.45*	–.61*	.67*	.59*	.65*	–.49*
Cantril Ladder of Life (now)		–.63*	–.64*	.50*	–.51*	.56*	.57*	.74*	–.52*
Center for Epidemiologic Studies—Depression Scale			.85*	–.45*	.73*	–.51*	–.50*	–.64*	.48*
Shacham Profile of Mood States—total				–.42*	.80*	–.47*	–.44*	–.58*	.51*
Bradburn Positive Affect Scale					–.35†	.38*	.23†	.40†	–.17
Bradburn Negative Affect Scale						–.35*	–.30*	–.42*	.42*
Self-rated Karnofsky Performance Scale							.74*	.67*	–.50*
Medical Outcomes Study SF-20 Physical Functioning Scale								.63*	–.59*
Medical Outcomes Study SF-20 General Health Perceptions Scale									–.58*

*Significance, $P = .01$.
†Significance, $P = .05$.

quires examining the pattern of correlations for the CES-D Scale, the POMS total score, and the Bradburn Negative Affect Scale. As shown in Tables 2 and 3, the CES-D and the POMS scales have their highest correlations with each other. In addition, the Bradburn Negative Affect Scale has its highest correlations with these two other measures of negative mood. All these scales have considerably smaller correlations with the Bradburn Positive Affect Scale, which is appropriate, as the two Bradburn scales are intended to be (approximately) independent of each other. As these tables show, that is not quite the case here, but the correlation of positive affect and negative affect is smaller than most of the correlations that each has with other QOL measures.

The measures that deal with physical functioning also show a pattern of higher intercorrelations than with other QOL measures. The SR-KPS measure of physical functioning and the MOS SF-20 Physical Functioning Scale show their highest correlations with each other. The lowest absolute values for correlations with the SR-KPS measure are the Bradburn Positive Affect and Negative Affect Scales. This is also true of the MOS SF-20 Physical Functioning Scale.

Current pain shows moderate correlations with nearly every measure, with the exception of weak or nonsignificant correlation with the Bradburn Positive Affect Scale. For the Hispanic sample, its strongest correlations are with the MOS SF-20 Physical Functioning Scale and the MOS SF-20 General Health Perceptions Scale. In the African-American sample, the strongest correlations are with the SR-KPS scale and the MOS SF-20 Physical Functioning Scale.

The MOS SF-20 General Health Perceptions Scale has its strongest correlations in the Hispanic sample with the Cantril Ladder of Life evaluating one's current life ($r = .74$). Close behind are the SR-KPS ($r = .67$) and the MOS SF-20 Physical Functioning Scale ($r = .63$), as well as the SLDS Scale ($r = .65$) and the CES-D Scale ($r = -.64$). In the African-American sample, the MOS SF-20 General Health Perceptions Scale has its strongest correlations with the MOS SF-20 Physical Functioning Scale ($r = .71$) and the SR-KPS ($r = .65$). The correlation of the MOS SF-20 General Health Perceptions Scale with the Bradburn Positive Affect Scale is notably stronger among Hispanics ($r = .40$, $P \leq .05$) than among African-Americans ($r = .14$, not significant).

The validity of measures for particular populations requires more than a single study. However, the initial findings are at least encouraging that there is good support for both convergent and discriminant validity for those broad QOL dimensions where several concordant measures are available for comparison.

Advantages of Graphic Response Formats

In facilitating response by respondents who have limited literacy levels, it is helpful to use pictures, cartoons, or other graphic response formats that minimize dependence on ability to read text. Our experience with the SLDS-C with its use of the "smiley-face" response format and with the Cantril Ladders of Life with their use of a ladder format has shown the advantages of using response formats that do not depend much on language facility.

The Dartmouth COOP Charts provide another example of an approach that uses pictures as a basis for eliciting information from patients. The COOP Chart system includes nine charts to screen and monitor patients' functioning and was specifically designed for use in physicians' offices during the doctor–patient interview (*62,63*). Stick-figure drawings are used in these charts to indicate different levels of functioning. C. C. Gotay (personal communication, 1995) and colleagues are developing the COOP Charts for assessment of QOL with several Asian-American and Hawaiian populations. They have developed new charts to answer additional demands and have made minor modifications to the COOP Charts on the basis of their pretesting.

Conclusions

The results of this analysis of our initial first-phase study offer encouragement regarding the potential for some available measures of different QOL dimensions to be used as culturally equivalent measures across African-American and Hispanic patient populations. Ten measures have been identified that appear to be worthy of further reliability and validity studies with culturally diverse cancer patient samples and eventual testing in multiple clinical trials.

One particular QOL dimension has been identified that requires special attention because of differences in values and normative assumptions across culturally diverse groups. That is the QOL dimension of family functioning; one of the most popular measures of family functioning, the FACES III, apparently failed to be acceptable because its middle-class, Eurocentric assumptions about the role of children in family decision making seemed too foreign to be taken seriously. Attaining culturally equivalent measures in the domains of family functioning offers a particularly interesting challenge for further research.

Finally, the procedures described here for assessing the cultural equivalence of QOL measures have been shown to be practical and productive. Techniques to establish cultural equivalence, utilization of special advisory boards, and completion of interviews in the community offer methods to evaluate available measures of QOL in diverse populations. These techniques are equally applicable to measures other than QOL and for use in other special populations.

References

(*1*) Holland JC, Silverfarb P, Tross S, Cella D. The Cancer and Leukemia Group B (CALGB) experience. In: Ventrafridda V, Van Dam FS, Yancik R, Tamburini M, editors. Assessment of quality of life and cancer treatment. Amsterdam: Excerpta Medica, 1986.

(*2*) Aaronson NK, Beckmann J. The quality of life of cancer patients. New York: Raven Press, 1987.

(*3*) Fayers PM, Jones DR. Measuring and analysing quality of life in cancer clinical trials: a review. Stat Med 1983;2:429-46.

(*4*) Holland JC. Methodology in behavioral and psychosocial cancer research. Need for improved psychosocial research methodology: goals and potentials. Cancer 1984;53(10 Suppl):2218-20.

(*5*) de Haes JC, van Knippenberg FC. The quality of life of cancer patients: a review of the literature. Soc Sci Med 1985;20:809-17.

(*6*) Ventafridda V, Van Dam FS, Yancik R, Tamburini M, editors. Assessment of quality of life and cancer treatment. Amsterdam: Excerpta Medica, 1986.

(*7*) Clark A, Fallowfield LJ. Quality of life measurements in patients with malignant disease: a review. J R Soc Med 1986;79:165-9.

(*8*) Selby P, Robertson B. Measurement of quality of life in patients with cancer. Cancer Surv 1987;6:521-43.
(*9*) Donovan K, Sanson-Fisher RW, Redman S. Measuring quality of life in cancer patients. J Clin Oncol 1989;7:959-68.
(*10*) Aaronson NK, Meyerowitz BE, Bard M, Bloom JR, Fawzy FI, Feldstein M, et al. Quality of life research in oncology. Past achievements and future priorities. Cancer 1991;67(3 Suppl):839-43.
(*11*) Moinpour CM, Savage M, Hayden KA, Sawyers J, Upshurch C. Quality of life assessment in cancer clinical trials. In: Dimsdale JE, Baum A, editors. Quality of life in behavioral medicine research. Hillsdale (NJ): Lawrence Erlbaum Associates, 1995:79-95.
(*12*) Payne KW, Ugarte CA. The Office of Minority Health Resource Center: impacting on health related disparities among minority populations. Health Educ 1989;20:6-8.
(*13*) Baquet CR, Hunter CP. Patterns in minorities and special populations. In: Greenwald P, Kramer BS, Weed DL, editors. Cancer prevention and control. New York: Marcel Dekker, 1995:23-36.
(*14*) Anastasi A. Psychological testing. 5th ed. New York: Macmillan, 1982.
(*15*) Ponterotto JG. Racial/ethnic minority research in the *Journal of Counseling Psychology*: a content analysis and methodological critique. J Counseling Psychol 1988;35:410-8.
(*16*) Casas JM. Policy, training and research in counseling psychology: the racial/ethnic minority perspective. In: Brown S, Lent R, editors. Handbook of counseling psychology. New York: Wiley, 1984:785-831.
(*17*) Lonner RP, Ibrahim FA. Assessment in cross-cultural counseling. In: Pedersen PB, Draguns JG, Lonner WJ, Trimble JE, editors. Handbook of psychotherapy and behavior change: an empirical analysis. 2nd ed. New York: Wiley, 1989:903-38.
(*18*) Sue DW. Effective multicultural counseling: proposed research directions. In: Fouad NA, Chair. Cross-cultural research and training: focus on critical issues. Symposium presented at the annual meeting of the American Psychological Association, New Orleans, August 1989.
(*19*) Sue S. Psychotherapeutic services for ethnic minorities. Two decades of research findings. Am Psychol 1988;43:301-8.
(*20*) Flaherty JA, Gaviria FM, Pathak D, Mitchell T, Wintrob R, Richman JA, et al. Developing instruments for cross-cultural psychiatric research. J Nerv Ment Dis 1988;176:257-63.
(*21*) Choi IC, Comstock GW. Interviewer effect on responses to a questionnaire relating to mood. Am J Epidemiol 1975;101:84-92.
(*22*) Baker F, Intagliata J. Quality of life in the evaluation of community support systems: a review of the literature. Eval Program Planning 1982;5:69-79.
(*23*) Baker F, Jodrey D, Intagliata J. Social support and quality of life of community support clients. Community Ment Health J 1992;28:397-411.
(*24*) Andrews FR, Withey SB. Social indicators of well-being: Americans' perceptions of life quality. New York: Plenum Press, 1976.
(*25*) Baker F, Curbow B, Wingard J. Development of the Satisfaction With Life Domains Scale for Cancer (SLDS-C). J Psychosocial Oncol 1992;10:75-90.
(*26*) Baker F, Curbow B, Wingard JR. Role retention and quality of life of bone marrow transplant survivors. Soc Sci Med 1991;32:697-704.
(*27*) Cantril H. The pattern of human concerns. New Brunswick (NJ): Rutgers Univ Press, 1965.
(*28*) Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Measurement 1977;1:385-401.
(*29*) Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J Psychiatry 1980;137:1081-4.
(*30*) Roberts RE, Vernon SW. The Center for Epidemiologic Studies Depression Scale: its use in a community sample. Am J Psychiatry 1983;140:41-6.
(*31*) Naughton MJ, Wiklund I. A critical review of dimension-specific measures of health-related quality of life in cross-cultural research. Qual Life Res 1993;2:397-432.
(*32*) Roberts RE. Reliability of the CES-D Scale in different ethnic contexts. Psychiatry Res 1980;2:125-34.
(*33*) Zonderman AB, Costa PT Jr, McCrae RR. Depression as a risk for cancer morbidity and mortality in a nationally representative sample [*see* comment citation in Medline]. JAMA 1989;262:1191-5.
(*34*) Gritz ER, Wellisch DR, Siau J, Wang HJ. Long-term effects of testicular cancer on marital relationships. Psychosomatics 1990;31:301-12.
(*35*) Devins GM, Orme CM, Costello CG, Binik YM. Measuring depressive symptoms in illness populations. Psychometric properties of the Center for Epidemiologic Studies Depression (CES-D) Scale. Psychol Health 1988;2:139-156.
(*36*) Fobair P, Hoppe RT, Bloom J, Cox R, Varghese A, Spiegel D. Psychosocial problems among survivors of Hodgkin's disease. J Clin Oncol 1986;4:805-14.
(*37*) McNair DM, Lorr M, Droppleman LF. EITS manual for the Profile of Mood States. San Diego (CA): Educational and Industrial Testing Service, 1971.
(*38*) Eichman WJ. Review of the Profile of Mood States. In: Buros OK, editor. The eighth mental measurements yearbook. Highland Park (NJ): Gryphon Press, 1978.
(*39*) Shacham S. A shortened version of the Profile of Mood States. J Pers Assess 1983;47:305-6.
(*40*) Achterberg J, Lawlis GF. A canonical analysis of blood chemistry variables related to psychological measures of cancer patients. Multivariate Exp Clin Res 1979;4:1-10.
(*41*) McCorkle R, Quint-Benoliel J. Symptom distress, current concerns and mood disturbance after diagnosis of life threatening disease. Sci Med 1984;17:431-8.
(*42*) Taylor SE, Lichtman RR, Wood JV. Attributions, beliefs about control, and adjustment to breast cancer. J Pers Soc Psychol 1984;46:489-502.
(*43*) Worden JW, Weisman AD. Preventative psychosocial intervention with newly diagnosed cancer patients. Gen Hosp Psychiatry 1984;6:243-9.
(*44*) Cassileth BR, Lusk EJ, Walsh WP, Doyle B, Maier M. The satisfaction and psychosocial status of patients during treatment for cancer. J Psychosoc Oncol 1989;7:47-57.
(*45*) Cassileth BR, Lusk EJ, Strouse TB, Miller DS, Brown LL, Cross PA. A psychological analysis of cancer patients and their next-of-kin. Cancer 1985;55:72-6.
(*46*) Giacquinta B. Helping families face the crisis of cancer. Am J Nurs 1977;77:1585-8.
(*47*) Huth CM. Illness and the family. Ann Intern Med 1978;89:132-3.
(*48*) Plumb MM, Holland J. Comparative studies of psychological function in patients with advanced cancer—I. Self-reported depressive symptoms. Psychosom Med 1977;39:264-76.
(*49*) Graydon JE. Factors that predict patients' functioning following treatment for cancer. Int J Nurs Stud 1988;25:117-24.
(*50*) Bradburn NM. The structure of psychological well-being. Chicago: Aldine, 1969.
(*51*) Beiser M. Components and correlates of mental well-being. J Health Soc Behav 1974;15:320-7.
(*52*) Coward DD. Self-transcendence and emotional well-being in women with advanced breast cancer. Oncol Nurs Forum 1991;18:857-63.
(*53*) Wingard JR, Curbow B, Baker F, Piantadosi S. Health, functional status, and employment of adult survivors of bone marrow transplantation. Ann Intern Med 1991;114:113-8.
(*54*) Stewart AL, Hayes RD, Ware JE. The MOS short-form general health survey: reliability and validity in a patient population. Unpublished working draft. Santa Monica (CA): The RAND Corporation, 1987.
(*55*) Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
(*56*) Helmstadtler GC. Principles of psychological measurement. New York: Appleton-Century-Crofts, 1964.
(*57*) Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill, 1978.
(*58*) Zabora J, Smith ED, Baker F, Wingard J, Curbow C. The family: the other side of bone marrow transplantation. J Psychosoc Oncol 1992;10:35-46.
(*59*) Olson DH, Portner J, Lavee Y. FACES III. St. Paul (MN): Family Social Science, Univ Minnesota, 1985.
(*60*) Olson DH. Circumplex Model VII: validation studies and FACES III. Fam Process 1986;25:337-51.
(*61*) Crowne D, Marlowe D. A new scale of social desirability independent of psychopathology. J Consulting Psychol 1960;24:349-54.
(*62*) Nelson E, Wasson J, Kirk J, Keller A, Clark D, Detrich A, et al. Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. J Chronic Dis 1987;40 Suppl1:55S-69S.
(*63*) Nelson EC, Landgraf JM, Hays RD, Wasson JH, Kirk JW. The functional status of patients. How can it be measured in physicians' offices? Med Care 1990;28:1111-26.

Notes

Supported by Public Health Service grant 1R01CA61673 from the Nationa Cancer Institute, National Institutes of Health, Department of Health and Humar Services.

We acknowledge and give our special thanks to the members of the two community advisory boards who have participated in the instrument review proces described in this article. The members of the African American Advisory Boarc included Ms. Nina Harper, Dr. Miles Harrison, Dr. Melva Owens, Pastor Marshall Prentice, Ms. Demetria Rogers, and Rev. Melvin Tuggle II. The member of the Hispanic Advisory Board included Sister Mary Neil Corcoran, Ms. Libby Garcia-Herd, Dr. Elmer Huerta, Mr. Pedro Ponceano, Ms. Haydee Rodriguez and Dr. Ruth Zambrana.

Quality-of-Life Research in the Pediatric Oncology Group: 1991-1995

*Andrew S. Bradlyn, Brad H. Pollock**

Quality-of-life end points for cancer clinical trials have received much attention in the adult literature. However, within pediatric cancer clinical trials, the inclusion of these alternate end points has only recently been considered. We review the Pediatric Oncology Group's approach to research in this area and describe our guidelines and protocols that incorporate quality-of-life end points and several of the methodologic barriers that must be addressed. [Monogr Natl Cancer Inst 1996;20:49-53]

Over the past 10 years, there has been increasing emphasis on the role of alternate end points, such as quality of life (QOL), in cancer clinical trials. For example, QOL data have been used in a variety of phase III clinical trials that have compared different treatment modalities (*1*), regimens hypothesized to reduce toxicity (*2*), and the long-term psychosocial adjustment of survivors (*3*). Additionally, QOL data have been examined as prognostic factors in attempts to identify patient characteristics associated with differential responsiveness to therapeutic regimens (*4*). However, QOL end points have not typically been incorporated into pediatric cancer clinical trials. A recent retrospective review of phase III pediatric trials reported that less than 5% included QOL outcomes by almost any definition (*5*). Since 1991, the Pediatric Oncology Group (POG) has made a concerted effort to examine the potential contribution of end points, such as QOL, in the evaluation of alternate therapies. These end points are currently being incorporated in a variety of clinical trials.

Background

In 1980, the pediatric sections of the Cancer and Leukemia Group B and the Southwest Oncology Group were merged to form the POG. The two National Cancer Institute-sponsored pediatric cooperative groups (POG and the Children's Cancer Group) provide protocol-driven treatment for the vast majority of children and adolescents in the United States who are diagnosed with a malignancy (*6*). As a result of continued advances in diagnosis and treatment, the prognosis for many of these patients has changed dramatically over the past 30 years. For example, while the 5-year survival rate for patients diagnosed with acute lymphoblastic leukemia in the early 1960s was less than 5%, more recent data project a greater than 70% 5-year survival rate for these patients (*7*). However, while our effectiveness in improving the quantity of children's survival has been substantial, our understanding of the quality of their survival has been limited.

Approximately 4 years ago, the POG established a formal mechanism for examining the role that QOL end points might play in its clinical trials. At the direction of the Supportive Care and Cancer Control Committee, a subcommittee on QOL was established and convened its first meeting at the fall 1991 group meeting.

Accomplishments of the Committee

The committee comprises representatives from a variety of disciplines, including pediatric oncology, psychology, nursing, data management, epidemiology, and biostatistics. Initially, it was agreed that we needed to establish a consensus definition of QOL in pediatric oncology and to provide written guidelines to steer the POG's efforts in this area. Therefore, we developed a set of guidelines (reviewed below) that define QOL, discuss the selection and timing of measures, and address the quality control of this data collection. The objectives of these guidelines were to arrive at an organized, uniform, and coherent framework for QOL assessment within the POG and to employ methodologically sound data collection procedures and instruments to address these research questions. Additionally, we wished to answer QOL questions in an efficient manner, maximizing the benefits in relation to their cost, in terms of both financial and human capital.

Definition

We adopted the following definition of QOL, based in large part on the World Health Organization's definition of health (*8*):

> Quality of life is a multidimensional construct, incorporating both objective and subjective data, including (but not limited to) the social, physical, and emotional functioning of the child and, when indicated, his/her family. QOL measurement must be sensitive to changes that occur throughout development [POG; unpublished manuscript, 1993].

**Affiliations of authors:* A. S. Bradlyn, Department of Behavioral Medicine and Psychiatry, Robert C. Byrd Health Sciences Center, West Virginia University, Morgantown; B. H. Pollock, Department of Health Policy and Epidemiology, University of Florida College of Medicine, Pediatric Oncology Group Statistical Office, Gainesville.

Correspondence to: Andrew S. Bradlyn, Ph.D., Department of Behavioral Medicine and Psychiatry, Robert C. Byrd Health Sciences Center, West Virginia University, 930 Chestnut Ridge Rd., Morgantown, WV 26505.

This definition is consistent with those of other investigators (*9,10*) and was intended to provide a focus for our research efforts, so that they would be planned, coordinated, and responsive to the rigors of the scientific method. It highlights the unique importance of the role of child development in our patients and their response to treatment, the impact of treatment on development, and the impact on family members.

Prioritization of Clinical Trials for QOL Assessment

Because of the drain on limited resources imposed by the collection of these types of data, a number of factors were specified to aid in the identification of clinical trials in which QOL end points would be most relevant. This serves the purpose of prioritizing which trials would include an assessment of QOL. Consistent with the results of other investigators (*11*), randomized phase III trials are identified as being the most relevant setting for inclusion. While phase I and II trials are noted to have incorporated potentially relevant QOL questions, single-arm trials and those without randomization are considered to be problematic from a QOL research standpoint. Trials that are expected to accrue a sufficient number of patients for statistical power considerations are also identified as potentially promising. The guidelines note that QOL end points should be considered in trials of therapeutic equivalence, in comparative investigations of new treatment modalities (or comparisons of alternate treatment modalities), in those protocols with a supportive care or rehabilitation focus and in those that are likely to have a major impact on clinical practice.

Instrumentation

One of the critical responsibilities of the committee has been the identification of instruments that are consistent with our conceptual definition of QOL and protocol-specific research goals. This has been a dynamic process where the performance of instruments is reviewed on a regular basis; newly developed and published instruments are also critically reviewed for possible inclusion. A guiding principle of the committee is that QOL measurement must reflect current scientific standards for questionnaire development, psychometric properties, and practical use.

We elected to adopt a core set of QOL measures, including generic and cancer-specific measures. Investigators were free to include additional, specific modules (e.g., treatment or disease specific) to these core instruments for new protocols. We felt that recommending a consistent, core set of instruments would provide for the greatest degree of comparability of data across protocols, would increase quality control in a multi-institutional setting by reducing the number of different data forms, and would also enable us to continually update our knowledge and understanding of both the instruments and children's QOL. Unfortunately, the state of the art in measuring child and adolescent QOL is significantly behind that of adult QOL, where there are numerous instruments for both generic (e.g., SF-36; *12*) and cancer specific (e.g., Functional Assessment of Cancer Therapy; *13*) assessment. Based on our review of the literature, we recommended the following core instruments:

1. *Generic health status:* The instruments derived from the Rand Health Insurance Experiment (*14,15*) were identified as being the most appropriate, currently available instruments for our patient population. A major logistical barrier to the inclusion of QOL end points has been the age range that our patients represent (from birth to over 21 years of age for most malignancies, and up to 30 years of age for bone tumor protocols) and the lack of a single instrument that could cover that great an age span. It was recognized early in our discussions that we would need to be able to collect data across more than one developmental stage (e.g., 0-4 and 5-13 years of age) and that some degree of conceptual consistency was important in our choice of instruments. The Rand scales (0-4, 5-13, and 14 years of age and above) were therefore recommended because of their multidimensional focus, sound development and standardization, and desirable psychometric properties. Other scales, such as the Quality of Well-Being Scale (*16*), were considered but not recommended because of the nature of their administration (most typically administered by a trainer interviewer) and the associated cost of those procedures in a multi-institutional setting.

2. *Cancer-specific instruments:* One frequently noted limitation of generic health status instruments is the potential lack of sensitivity to issues that might be of great importance in a particular population (*17*). Cancer-specific instruments are recommended for inclusion to maximize the probability that the most relevant information will be obtained. Unfortunately, there is a paucity of such instruments currently available, but we have focused on two: the Pediatric Oncology Quality of Life Scale (*9*) and the University of Miami Quality of Life Scale (Armstrong FD; unpublished manuscript, 1995), both of which were developed and standardized with pediatric oncology patients and their families. Both of these instruments are completed by parents, and investigators are directed to choose one for inclusion in the study.

3. *Other recommended core assessments:* A variety of other, brief instruments are recommended in the core battery. These include the Play Performance Scale for children (*18*), which is a 0-100 parents-completed rating of their child's activity over the past 2 weeks (similar in form to the Karnofsky Performance Status scale for adults), and two single-item ratings. The first of these ratings is one of overall QOL (from the parents' or patient's perspective), and the second asks for a rating of overall health. These ratings are included to provide an additional, categorical evaluation of the patients that may be used in subsequent analyses. For adult survivors, the Karnofsky Performance Status Scale (*19*) is substituted for the Play Performance scale.

A number of measurement issues cloud the assessment of QOL for pediatric patients. As mentioned previously, it is not unusual for therapeutic trials to cover a wide age range (birth to 21 years) and to include a substantial follow-up period while on study, which may range up to 5 years. To highlight the many decision points, consider a child who initially registers in a protocol at age 4 years and is then followed for 5 years. In terms of QOL assessment, the parents would complete the 0–4-year-old version of the Rand scales and the single-item ratings. However, when the child is reassessed at age 6 years and beyond, should the parents complete the 0-4 Rand or the more age-appropriate version (5-13 years)? While there is item content overlap between the 0-4 and 5-13 versions, the earlier version has an

emphasis on acquisition of developmental milestones, while the scale for older children includes mental health items and school and social functioning as well. Also consider the situation in which the child is placed in the study at age 13 years; his or her parents complete the 5–13-year-old Rand and the other recommended core instruments. However, at age 14 years, the patient could potentially serve as his or her own informant and complete the 14 and above version of the Rand. A change in instruments at this point would introduce two potential confounds: instrument (5-13 versus 14 Rand) and informant (parent versus patient).

Generally speaking, our guidelines recommend that situations such as these be dealt with by having the same instrument completed at each data collection point. For all intents and purposes then, this means that whatever instrument was completed at registration should be completed at each subsequent assessment, recognizing that there will be a group of patients who are "out of bounds" for the instrument. However, given the emphasis on measuring change over time, maintaining the consistency of the instrument is a higher priority.

An additional and troubling issue is the use of proxy respondents (i.e., parents) in pediatric trials. There is widespread acceptance of the patient as the best provider of QOL data (*20*), but pediatric patients provide for many exceptions to this maxim. As noted above, these trials may include patients not judged competent to provide this information as a function of their age (e.g., 3-year-olds). The lower age at which children can competently provide these types of data is not well established, but there is a recent report of pediatric cancer patients as young as 5 years of age being able to do so in a reliable and valid manner (Kaplan S, Barlow S, Spetter D, Sullivan L, Khan A, Parsons S, et al; unpublished manuscript, 1995). Proxy respondents are also necessary in those situations for which child or adolescent self-report measures are not available. For example, at this time there are no published self-report cancer-specific instruments for school-aged children and adolescents or for the generic health status assessment of younger, school-aged patients. The agreement between patient and proxy measures in this population is not well understood and may, in fact, be quite poor (Kaplan et al.; unpublished manuscript, 1995). The POG QOL guidelines recommend that to the extent that it is feasible, the patient's perspective should be included. Toxicity measures, in and of themselves, are not appropriate as proxy measures, since they include little if any assessment of the impact of those toxic effects on the patient's functioning.

Other Recommendations Included in the Guidelines

There are several other issues that merit brief dicussion. At the time the guidelines were initially developed, it was typical for QOL questions to be addressed in a companion protocol to a therapeutic protocol. At that same time, however, there was discussion about the difficulties arising from such a strategy, the primary one being the situation where only a small percentage of patients were registered on the QOL study in comparison to those actually receiving treatment. This introduced a number of difficulties into the interpretation of the data, not the least of which was potential selection bias. Our guidelines now recommend that QOL questions be included in the therapeutic protocol and that they be labeled as specific cancer control objectives. This has the benefit of allowing for joint review by both the treatment (Cancer Therapy Evaluation Program) and cancer control (Division of Cancer Prevention and Control) divisions of the National Cancer Institute, with assignment of cancer control credits (that provide financial support for institutional data management).

Protocols with a QOL component should also have an individual identified as the coordinator for those particular research questions. This person is responsible for the design and implementation of the data collection strategy, as well as the ongoing monitoring of the data quality. Each participating institution has also been asked to identify an individual on site who will be responsible for ensuring that patients complete the required QOL assessments at the specified times and in the specified manner. A manual was recently developed and distributed to these individuals to maximize the quality and quantity of data collection.

Protocols With QOL End Points

A number of phase III trials that are currently accruing patients or are expected to begin accrual in the near future include QOL end points. Several earlier trials included a modified or shortened version of our core battery, most typically bundled with a series of psychologic or neuropsychologic evaluations. Protocols that include the core battery, as recommended by POG include:

1. *POG 9485/9585: "Evaluation of the Role of Minimal Access Surgery in the Treatment of Childhood Cancers."* For this pair of intergroup studies (with the Children's Cancer Group), the objectives are to evaluate the role of minimal access surgery versus standard open approach surgery in the diagnosis, staging, and treatment of a wide range of tumors. The primary end points are treatment-related morbidity and mortality, QOL, and economic costs. QOL assessments are scheduled at base line (registration), postsurgical day 3, and postsurgical day 30, allowing for examination of the differential effects of these surgical procedures on short-term or acute QOL.

2. *POG 9421: "Evaluation of Standard vs. High-Dose Ara-C Induction Followed by the Randomized Use of Cyclosporine as an MDR Reversal Agent, Compared to Allogeneic BMT in Childhood AML (Phase III)."* The objective of this study relating to QOL is to compare the event-free survival between patients randomized between allogeneic bone marrow transplantation and chemotherapy. Serial QOL assessments and neuropsychologic evaluations are scheduled to identify the effects of the two treatments on the patient's QOL. This study is projected to accrue 150 patients per year, for each of 4 years.

3. *POG 9315: "A Phase III Study of Large-Cell Lymphomas in Children and Adolescents: Comparison of Cytarabine, Prednisone, and Vincristine Versus Cytarabine, Prednisone, and Vincristine Plus Methotrexate, Hydroxyurea, and Cytarabine and Continuous Versus Bolus Infusion of Doxorubicin."* The primary objective of this investigation is to study whether intermediate-dose methotrexate and high-dose cytarabine administered during the maintenance phase can improve the event-free survival of patients with advanced-stage large cell

lymphoma. The objective relating directly to QOL focuses on the impact of doxorubicin infusion time (continuous versus bolus) on subsequent efficacy, cardiotoxicity, and QOL. This protocol is projected to accrue approximately 240 patients over a 5-year period.

4. *POG 9480: "Afterload Reduction for Late Anthracycline Cardiotoxicity."* The primary objective of this placebo-controlled, double-blind, randomized intervention trial is to investigate the role of enalapril in ameliorating the late cardiotoxic effects of doxorubicin for long-term survivors of childhood cancer. One of the specific objectives is to determine the impact of enalapril therapy on QOL. Additionally, we will compare, using the Q-TWiST method (quality-adjusted time without symptoms and toxicity), the tradeoff between treatment impact on QOL and cardiac functioning. This study is projected to accrue a total of 150 subjects who will be followed for a period of 4 years.

Future Goals and Directions

As part of our overall cancer control objectives and activities, we have come a long way toward including alternate end points in therapeutic protocols. However, as in any other developing area of research, we anticipate significant advances in our knowledge over the next several years. As interest in QOL assessment with children has burgeoned, new instruments are likely to become available and we will have gained further information regarding the performance of instruments in our core battery. Thus, the core set of instruments that is currently recommended is likely to change as more experience is gained in the group.

Our QOL efforts include providing ongoing education and consultation to the cooperative group membership in general. We have planned a variety of educational activities, including roundtable discussions at upcoming group meetings and continuing presentations and discussions in disease committees as new protocols are developed. It is a primary goal that protocols with relevant QOL questions be identified in the conceptualization and design phases so that these end points can be well integrated in protocols from the beginning and not included as a post hoc consideration. Early incorporation allows for the optimal integration of the QOL or cancer control objectives in the protocol and provides pivotal opportunities for full discussion regarding the hypothesized effects on QOL and how and when they may best be measured.

It is important to note that in many of our protocols, QOL end points do not stand alone and may be incorporated into a larger package of alternate end points, including factors such as economic analyses and conventional end points such as survival. Protocols with multiple dependent variables, such as survival, QOL, and cost, will generate datasets that provide the opportunity for exploration of the interrelationships among these end points, their relationship to independent variables, and the roles that they play in ultimate patient outcome. Additionally, as described in the enalapril preventive investigation, we are exploring the applicability of utility assessments, such as the Q-TWiST methodology, as a means to integrate quantity and quality of survival.

A final, and ultimately critical, goal is to understand how these data should be interpreted and how they will be integrated into clinical decision making. The decision-making process is straightforward in situations where survival rates between two groups are equal but there are differences in QOL or where the survival and QOL of patients in one group are both superior to that in the other group. The most problematic situation, however, is when one therapeutic arm has poorer survival but better QOL or when QOL improves but quality deteriorates.

Investigation of QOL issues in pediatric oncology (as compared to adult oncology) presents a number of unique situations in terms of integrating these data into clinical decision making. For example, we have previously noted (*21*) that inclusion of QOL end points in clinical trials has been hampered by the attitude that therapy for pediatric patients is curative in intent, while palliation may more often be the focus in adult clinical trials. Thus, there may be a willingness to trade quality of life for quantity of life in certain pediatric settings. How "treatment intent" should or can be integrated into decision making is unknown at this time and is an area for future investigation. Additionally, given the relatively positive prognosis for many pediatric malignancies, investigators must also examine the long-term or late effects of treatment and disease on QOL. While there are numerous reports of late effects on organ functioning, few if any have examined the actual impact of treatment on functioning. QOL data have the potential to go beyond the description of toxicity or late effects to actually describe how those effects are related to the patient's day-to-day abilities in their social, emotional, and physical functioning. These are important issues that are only beginning to be explored.

Because of the relative rarity of pediatric cancers, QOL research with this population typically occurs in the context of a cooperative group, multi-institutional setting. Approximately 1.5 million adults in the United States are diagnosed with cancer each year, which is orders of magnitude larger than the approximately 8000-10 000 children and adolescents diagnosed with cancer each year. However, the number of years of life saved is substantial in the pediatric population, and as the number of survivors increases, the study of the quality of their survival will come more to the forefront and become predominant in the development of new therapies. Multi-institutional research investigations are costly, but the potential payoff for society, particularly in regard to the contributions that these survivors may make, is likely to be substantial.

References

(*1*) Sugarbaker PH, Barofsky I, Rosenberg SA, Gianola FJ. Quality of life assessment of patients in extremity sarcoma clinical trials. Surgery 1982; 9:17-23.

(*2*) Selby P, Robertson B. Measurement of quality of life in patients with cancer. Cancer Surv 1987;6:521-43.

(*3*) Kornblith AB, Anderson J, Cella DF, Tross S, Zuckerman E, Cherin E, et al. Hodgkin's disease survivors at increased risk for problems in psychosocial adaptation. The Cancer and Leukemia Group B. Cancer 1992;70:2214-24.

(*4*) Kaasa S, Mastekaasa A, Lund E. Prognostic factors for patients with inoperable non-small cell lung cancer, limited disease. The importance of patients' subjective experience of disease and psychosocial well-being. Radiother Oncol 1989;15:235-42.

(5) Bradlyn AS, Harris CV, Spieth LE. Quality of life assessment in pediatric oncology: a retrospective review of phase III reports. Social Science Med. In press.
(6) Pediatric Oncology Group. Progress against childhood cancer: the Pediatric Oncology Group experience. Pediatrics 1992;89:597-600.
(7) Boring CC, Squires TS, Tong T. Cancer statistics, 1991. CA Cancer J Clin 1991;41:19-36.
(8) World Health Organization. The first ten years of the World Health Organization. Geneva, 1958.
(9) Goodwin DA, Boggs SR, Graham-Pole J. Development and validation of the Pediatric Oncology Quality of Life Scale. Psycholog Assess 1994;6:321-8.
(10) Mulhern RK, Ochs J, Armstrong DF, Horowitz ME, Friedman AG, Copeland D, et al. Assessment of quality of life among pediatric patients with cancer. Psycholog Assess 1989;1:130-8.
(11) Ware JE, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
(12) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.
(13) Nayfield SG, Hailey BJ, McCabe M. Report of the Workshop on Quality of Life Research in Cancer Clinical Trials. Bethesda, MD: National Cancer Institute, 1990.
(14) Eisen M, Donald C, Ware JE Jr, Brook R. Conceptualization and measurement of health for children in the health insurance study. Santa Monica, CA: Rand Publication R-2313 HEW, 1980.
(15) Brook RH, Ware JE Jr, Davies-Avery A, et al. Conceptualization and measurement of health for adults in the Health Insurance Study. Volume 8: Overview. Santa Monica, CA: Rand Publication R-1987/8 HEW, 1979.
(16) Kaplan RB, Anderson JP. A General Health Policy Model: Update and applications. Health Serv Res 1988;23:203-35.
(17) Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989;27, S217-32.
(18) Lansky SB, List MA, Lansky LL, Ritter-Sterr C, Miller DR. The measurement of performance status in childhood cancer patients. Cancer 1987; 60:1651-6.
(19) Karnofsky D, Burchenal J. Clinical evaluation of chemotherapeutic agents in cancer. In: Macleod CM, editor. Evaluation of chemotherapeutic agents. New York: Columbia Press, 1949:191-205.
(20) Osoba D. Lessons learned from measuring health-related quality of life in oncology [*see* comment citation in Medline]. J Clin Oncol 1994;12:608-16.
(21) Bradlyn AS, Ritchey AK, Harris CV, Moore AK, O'Brien RT, Parsons SK, et al. Quality of life research in pediatric oncology: research methods and barriers. Cancer. In press.

Model for Quality-of-Life Research From the Cancer and Leukemia Group B: the Telephone Interview, Conceptual Approach to Measurement, and Theoretical Framework

*Alice B. Kornblith, Jimmie C. Holland**

Cancer and Leukemia Group B Quality-of-Life Research Experience: 1976-1995

Historical Background: 1976-1985

In 1976, the Cancer and Leukemia Group B (CALGB), the oldest cancer cooperative clinical trials group, established the Psycho-Oncology (originally Psychiatry) Committee as a part of its multimodality structure. This created the first opportunity to ask psychological, social, and quality-of-life (QOL) questions in large patient groups in which medical variables were controlled and biases of geographic location, investigator, and treatment were reduced. Thus, the CALGB has provided a setting for more definitive testing of psychosocial questions relevant to cancer patients than could be achieved in studies conducted from single institutions.

The Psycho-Oncology Committee began by assessing relevant psychosocial variables in specific clinical trials and later began to add QOL as one of the outcome variables assessed in clinical trials. A core of instruments was used to measure the impact of treatment on physical, psychological, and social functioning and to assess the role of psychosocial and demographic factors on survival (*1-3*). In the first 8 years, more than 1000 patients treated in eight CALGB protocols were studied with the core battery of measures, permitting comparison of patients' psychosocial function and QOL across disease site and treatment. This database provided the early studies of the impact of sociodemographic (education and income) (*4*) and psychological characteristics on survival, controlling for disease, treatment, and prognostic variables (*5,6*). Furthermore, it enabled the comparison of psychological distress of patients with advanced pancreatic cancer to those with advanced gastric cancer when receiving similar treatment protocols (*7*), the development of norms for the Profile of Mood States, a widely used measure of psychological distress, as well as the creation of a brief, valid version of the measure (*8,9*), and the examination of the relationship of disease stage and performance status to psychological state in patients with lung cancer (*10*).

By the early 1980s, the CALGB had accrued one of the largest cohorts of leukemia and Hodgkin's disease survivors who had been treated in CALGB phase III protocols. In 1985, the Psycho-Oncology Committee used this unusual resource to examine the long-term psychosocial adaptation of these survivors, using a core of instruments. The first studies explored the negative impact of cranial radiation in childhood leukemia (*11*), followed by examination of the QOL of adult-onset Hodgkin's disease survivors (*12-15*), leukemia survivors (*16*), and 15-year survivors of breast cancer (in progress).

Over the past 5 years, there has been an explosion of interest by the cooperative clinical trials groups in assessing QOL. As a consequence of many clinical trials detecting only marginal differences in survival among treatment arms, QOL outcomes have assumed greater importance. In 1985, partly in response to this issue, the Food and Drug Administration required that for new anticancer agents to gain approval for use, either a primary survival gain or a secondary QOL benefit must be demonstrated (*17*). Concurrently, there has been a rapid increase in the number, reliability, and validity of QOL measures with the "Handbook of Quality of Life Measures" (*18*), a compendium of QOL measures commonly employed in cancer research today, bearing testimony to this expansion.

The enhanced interest in QOL research, coupled with strong, stable, collaborative ties between CALGB psycho-oncology and oncology investigators, led the group to address the need to improve QOL data collection procedures. A major limitation of QOL research in cooperative clinical trials groups has been that the responsibility for data collection was placed on busy data managers and research nurses, without designated time or additional funding. Data often were not collected at the correct time, due to hectic clinic schedules. Data managers often did not have time to explain or instruct patients in the use of questionnaires, which frequently resulted in missing data and invalid scores. Patients, anxious prior to being seen by their oncologist or receiving treatment, provided ratings of their psychological state that may not have reflected their usual condition, resulting in skewed results; others were often reluctant to fill out questionnaires at all. As a consequence of these problems, there were frequent missing data points, typically 50% at base line, with substantial further attrition over the course of the study, raising serious concerns as to the representativeness of the sample,

**Affiliation of authors:* Memorial Sloan-Kettering Cancer Center, New York, NY.

Correspondence to: Alice B. Kornblith, Ph.D., Psychiatry Service, Box 266, Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021.

See "Note" section following "References."

validity, and generalizability of the findings. In 1985, in an effort to deal with these plaguing problems, the Psycho-Oncology Committee changed the data collection method to interviewing patients at home by telephone.

Data Collection by Telephone Interview

The decision to collect data by telephone interview was based in part on research over the past 25 years that had demonstrated the efficacy of telephone interviewing for the collection of psychosocial data. Because telephones are present in over 95% of American households (*19*), a largely representative sample could be captured using this method of data collection. For special segments of the population who could not be interviewed by telephone either because of lack of access to a phone or hearing problems, as is more likely with the socioeconomically disadvantaged and the elderly, respectively (*20*), a mixed-mode method of data collection (*21,22*) could be applied, using mailed questionnaires in place of telephone interviewing. Studies have shown that there are few substantive differences between results obtained from telephone and in-person interviewing, that the interrelationship among variables is maintained with both methods, and that the level of missing data is comparable (*23-30*). Further, less distortion in reporting socially undesirable acts occurs with the anonymity provided by the telephone interview (*25,29*). Response biases that have been reported to occur more frequently in telephone than in in-person interviewing are greater use of extreme ends of response categories (e.g., "never" and "extremely") and briefer responses to open-ended questions. For questions requesting ratings of agreement with given statements, there is a greater willingness to agree (i.e., acquiesce) with telephone interviewing, regardless of the statements' content (*26,28*). When costs incurred by the two methods have been analyzed, telephone interviewing has consistently been shown to be less expensive than in-person interviewing (*19,26,27,31,32*).

Hodgkin's disease survivor study. In 1986, telephone interviewing was initiated in a study of survivors of advanced Hodgkin's disease (CALGB 8561 and 8562) (*12-14*), which was supported by the CALGB budget from the National Cancer Institute for a trained telephone interviewer. Procedurally, patients were sent a letter from the Principal Investigator of the institution where they were treated that explained the study and informed them that a research interviewer would call within 1-2 weeks to discuss the study with them. Upon calling the patient, the interviewer answered questions concerning the study, obtained the patient's consent to participate, and made an appointment for the telephone interview within 7-10 days. A packet was mailed to the patient that contained two copies of the consent form (one of which was signed and returned), along with a copy of the psychosocial measures. The patient was instructed to read through the material and answer as many of the questions as possible prior to the interview. At the time of the telephone interview, the interviewer clarified questions the patient had and entered his/her answers on an identical form. The telephone interview usually lasted 45-60 minutes.

Of the 369 eligible patients, 273 (74%) survivors of Hodgkin's disease were interviewed. This 74% participation rate was high and well within the range of 49%-82% reported for telephone interviewing (*24,26-28,30,32*). The refusal rate was only 9%, near the lowest level in reported refusal rates of 4%-29% (*26,28,31*). An additional 15% were not interviewed because they were lost to follow-up and could not be located. Missing data were considerably diminished (*12*). The results of the CALGB study were compared with the study by Fobair et al. (*33*) of early and advanced Hodgkin's disease survivors whose data-collection method had been by in-person interviews and self-report questionnaires. On equivalent questions concerning problems patients' attributed to cancer, findings were remarkably similar between the CALGB (*12*) and Fobair et al. (*33*) studies: decrease in sexual activity (21% versus 20%, respectively), divorce or separation (56% versus 49%, respectively), and loss of job (5% versus 6%, respectively). This provided further evidence of the comparability of data obtained by telephone and in-person interview (*12*).

Importantly, it became evident that many patients enjoyed the interview. Some openly stated that the interview provided an opportunity to discuss a broad range of illness-related issues with an interested, caring individual. They found it gratifying that the CALGB had a continued interest in their well-being. While similar comments have been made by patients in active treatment in our other studies involving telephone interviews, they have been particularly frequent from cancer survivors who no longer maintain frequent contact with their oncologists.

Expansion of the telephone interview method to other CALGB studies. The use of the telephone interview for QOL studies for patients during treatment was first tested in a QOL study of stage IV breast cancer patients in a phase III dose–response trial of megestrol acetate (CALGB 8864) (*34*). Patients were interviewed by telephone three times over a 3-month period using a battery of measures significantly shortened from that of the Hodgkin's disease survivor study to accommodate their limitations due to advanced-stage disease. Only 4% refused to participate after consent had been given. While there was attrition over the 3-month period because of disease progression resulting in termination from the clinical trial (21%), sickness (3%), and death (2%), most patients who were able to participate were assessed successfully by telephone interview. The combined experience of the studies of Hodgkin's disease survivors and breast cancer patients on megestrol acetate has resulted in the telephone interview becoming the standard data collection method for CALGB QOL studies (*35*).

Attrition due to illness demonstrated in the breast cancer megestrol acetate study underscores the limitations placed on QOL research in patients with advanced stages of disease using *any* self-report methodology. Our QOL study of cachectic patients in the terminal stage of illness, treated with megestrol acetate to increase their appetite and weight (CALGB 8971), met with this overriding limitation also. Experience suggests that obtaining self-report data from patients either in the terminal phase of their disease or with a poor performance status are highly limited and inappropriately intrusive. In these patients, behavioral observations of patient functioning made by a family member, caregiver, or health professional must suffice.

Approach to Measurement

Our measurement approach has been similar to that of Aaronson et al. (*36*), with major QOL dimensions assessed by a core of in-

struments supplemented by measures specific to the disease site and treatment protocol. This approach has been applied to the development of *two* core sets of measures for two distinct patient populations: those in active treatment and cancer survivors. Most recently, individuals at high genetic risk for cancer have emerged as a new study population, requiring a *third* core set of measures to adequately assess relevant QOL issues (Table 1).

Assessment of QOL of patients in active treatment. Because treatment protocols address different sites of cancer at different disease stages, assessing QOL of patients in active treatment requires a broad spectrum of measures. In the late 1980s, the Functional Living Index-Cancer (FLIC) (*37*) was used in several breast and prostate cancer trials (CALGB 8864, 9181, and 9182) as the core QOL measure. However, the FLIC provided only an overall QOL total score, without subscale scores for different QOL dimensions. We therefore switched our core QOL measure for several lung cancer phase III trials [CALGB 8931 (*38*); CALGB 9033] to the EORTC Quality of Life Questionnaire (EORTC QLQ-C30) (*39*) and the Lung Cancer Module (EORTC QLQ-LC13) (*40*) when they became available. The EORTC measures were considered well suited for these studies because they had been originally developed in patients with lung cancer, had demonstrated reliability and validity on large samples, had subscale scores for the essential domains of QOL, and were brief, posing less respondent burden for very sick patients. To strengthen the social functioning component of the EORTC QLQ-C30 measure, the Duke–University of North Carolina Social Support Questionnaire (*41*) was appended. Because of its demonstrated strengths, the EORTC QLQ-C30 has also served as the core measure for other CALGB studies involving patients with breast cancer (CALGB 9364), myelodysplastic syndrome (CALGB 9221), and pleural effusions treated by talc thoracoscopy versus talc slurry (CALGB 9334). As new measures are developed, they are routinely reviewed for their potential use in our trials. For example, the Functional Assessment of Cancer Therapy Scale (FACT) (*42*), with its core QOL component supplemented by modules for nine disease sites and specific treatment regimens (e.g., bone marrow transplant), is being considered for use in studies currently in development.

For studies involving patients at an earlier stage of disease or with better performance status, expanded measurement has been possible, enabling assessment of additional variables or more in-depth measurement of important constructs. Because of the centrality of psychological state to understanding patients' QOL, the Mental Health Inventory (MHI) (*43,44*) has been frequently used to obtain an assessment of both positive and negative affect (CALGB 8864, 9221, and 9364). Particular QOL dimensions or related constructs have been assessed through the addition of the following measures: the McCorkle Symptom Distress Scale (*45,46*), to assess physical symptoms in several breast and prostate cancer clinical trials (CALGB 9066, 9181, and 9182); the Memorial Symptom Assessment Scale (*47*), for a colorectal cancer trial in development (CALGB 9481); the Wisconsin Brief Pain Inventory (*48,49*), for an extended assessment of pain in several prostate cancer clinical trials (CALGB 9181,

Table 1. Measures used in selected CALGB studies

		Measure*															
	Study No.	BSI	POMS	IES	Sexual problems	EMP/ INS problems	Cond N&V	PAIS	Socio dem	FLIC	MHI	Rand Func Limit	EORTC	MOS Social Support Survey	McCorkle Symptom Distress Scale	Additional measures
Active treatment																	
Breast	8864								X	X	X	X				X
	9066							X	X	X					X	
Lung	9033								X				X			X
	9334								X				X			X
Prostate	9181								X	X		X			X	X
	9182								X	X		X			X	X
Myelodysplastic syndrome	9221							X	X		X		X			
Other, cachexia	8971								X	X	X	X				X
Survivors																
Hodgkin's disease	8561 and 8562	X	X	X	X	X	X	X	X							
Leukemia	8963	X	X	X	X	X	X	X	X							X
Breast (proposed)		X		X	X	X	X	X	X					X		X
Hodgkin's disease telephone counseling (proposed)		X		X		X	X	X	X							X
High-risk screening (proposed)									X		X					X
Special issues																
Bereavement	9364								X		X		X	X		X
Hydrazine sulfate	8931								X				X			X

*BSI = Brief Symptom Inventory; POMS = Profile of Mood States; IES = Impact of Event Scale; EMP/INS = Employment and Insurance problems; Cond N&V = conditioned nausea and vomiting; PAIS = Psychosocial Adaptation to Illness Scale; Socio dem = sociodemographic characteristics; FLIC = Functional Living Index-Cancer; MHI = Mental Health Inventory; and Rand Func Limit = Rand Functional Limitations Scale (modified).

9182, and 9480); and the Body Image Subscale (*50*), to assess the effects of increasing weight on patients' body image in the breast cancer megestrol acetate trial (CALGB 8864).

Assessment of QOL of cancer survivors. A core set of measures can be applied as well in survivors of the commonality of issues across survivors of any neoplasm. In both the Hodgkin's disease (CALGB 8561 and 8562) and adult leukemia (CALGB 8963) studies and the proposed breast cancer survivor study (to begin in the fall of 1995), the following measures constituted the core. The Psychosocial Adaptation to Illness Scale (PAIS) (*51*) was used to assess the impact of having had cancer on survivors' current psychosocial and sexual functioning. Psychological state, both in general (Brief Symptom Inventory) (*52*) and specific to cancer (Impact of Event Scale) (*53*), was assessed in depth because of its central importance to adaptation. Measures were created to assess the continuing impact of cancer on survivors' lives, in terms of their employment, income, obtaining health and life insurance, sexual problems, and conditioned nausea and vomiting in response to treatment-related stimuli (*12,16*). Because sterility was an important issue for many patients with Hodgkin's disease who had been treated with alkylating agents, a detailed section of the questionnaire was devoted to assessing the prevalance of pregnancy outcomes, child deaths, and illnesses (*15*). By applying a core set of measures across survivor studies, as well as a set of disease-specific measures, there is an opportunity to compare the psychosocial adaptation of different groups of survivors and to speak meaningfully to issues that are important to each.

Assessment of individuals at high genetic risk. With the increasing development of DNA testing, tumor markers, and other presymptomatic cancer-screening methods, QOL issues relevant to this set of individuals at high genetic risk have assumed greater importance. Our initial study, in development, will evaluate the psychological consequences of an intensified screening program for relatives of patients with colon cancer. Theoretical models guiding measurement will include the Health Belief Model (*54,55*), related models (*56,57*), and the transtheoretical model of change (*58*). Outcome variables and measures of importance are psychological state (e.g., Mental Health Inventory), particularly general anxiety and anxiety specific to high-risk individuals (e.g., Kash's Breast Cancer Anxiety Scale, modified for colon cancer) and adherence to screening recommendations. To test the influence of mediating factors on an individual's adherence to screening recommendations, as suggested by the theoretical models, the following variables will be assessed: family history of cancer; history of compliance to colon cancer diagnostic testing; social support (e.g., MOS Social Support Survey) (*59*); and preventive health practices and health beliefs about cancer and screening, such as susceptibility to having cancer and potential costs and benefits of screening (e.g., General Health Motivation and Practices Scale) (*60*). The proposed study is the first in a series, as the CALGB develops a research program in the molecular genetics of solid tumors and hematopoietic malignancies. With the development of a CALGB registry of patients with breast cancer who consent to genetic testing, the psychological, ethical, and health behavior issues of genetic testing for both the patients and their relatives will become the focus of intensive study.

Special QOL research questions addressed by the CALGB. Some major psychological questions are ideally addressed in the cooperative group setting because of the large numbers of patients treated in clinical trials in which treatment is by protocol, with medical, treatment, and treatment-related outcome data stored in a common database. One area in which cooperative clinical trials groups are particularly valuable is the testing of the efficacy and QOL impact of an alternative therapy in a rigorously controlled trial. The CALGB conducted a phase III clinical trial of hydrazine sulfate, in which all patients with advanced non–small-cell lung cancer were entered in a randomized study to receive either a standard chemotherapy regimen and hydrazine sulfate or the chemotherapy regimen and placebo in a double-blind fashion (CALGB 8931) (*38*). The EORTC QLQ-30 (*39*) and Social Support LC13 (*40*) were used as the QOL measures for the study, supplemented by the Duke–University of North Carolina Questionnaire (*41*). No differences in survival or disease-free survival were found between the two groups, with the hydrazine arm of the study actually demonstrating worse physical functioning, greater fatigue, and worse lung cancer-specific symptoms than the control group at the 2-month assessment (*38*).

A second study currently under way, supported by the John D. and Catherine T. MacArthur Foundation, is quite different from others we have conducted in that it will test whether a major stressor, defined as loss of a spouse or child, is associated with a significantly increased risk of recurrence or death from breast cancer (CALGB 9364). The question of the relationship of stress to disease onset and recurrence is one that preoccupies many patient's concerns and is the focus of much research. However, studies of smaller cohorts have resulted in contradictory findings (*61*). By strictly defining the stressor in terms of what is universally accepted as a major stress, the loss of a spouse or child, a more definitive answer to this question may be provided. As a secondary objective to this study, the role of social support and spiritual beliefs in modulating the trauma of cancer will be examined. In this case–control study, in which all women were treated for stage II breast cancer over 10 years ago in CALGB 8541, case subjects were defined as women who had disease recurrence subsequent to their treatment in CALGB 8541 or who died; control subjects were those who were alive without disease progression. The odds of bereavement will be statistically compared between the two groups, with the odds ratio reflecting the increased risk of disease recurrence or death from breast cancer due to bereavement. The objectives led to the use of a psychosocial battery containing the MOS Social Support Survey to measure social activities and emotional and instrumental support from family and friends (*59*); the Life Experiences Survey (*62*) to assess stressful life events; the Systems of Belief Inventory to assess spirituality (Holland and Kash; personal communication); and the Texas Revised Inventory of Grief (*63*) to assess severity of bereavement for those who had experienced the death of a spouse or child.

Theoretical framework. In the past 5 years, the stress–illness vulnerability theory has emerged as a theoretical model for understanding patient adaptation (*64*). In this model, adapted for application to cancer patients (Fig. 1) (*65*), cancer and its treatment are the stressors, and QOL or patient adaptation is the out-

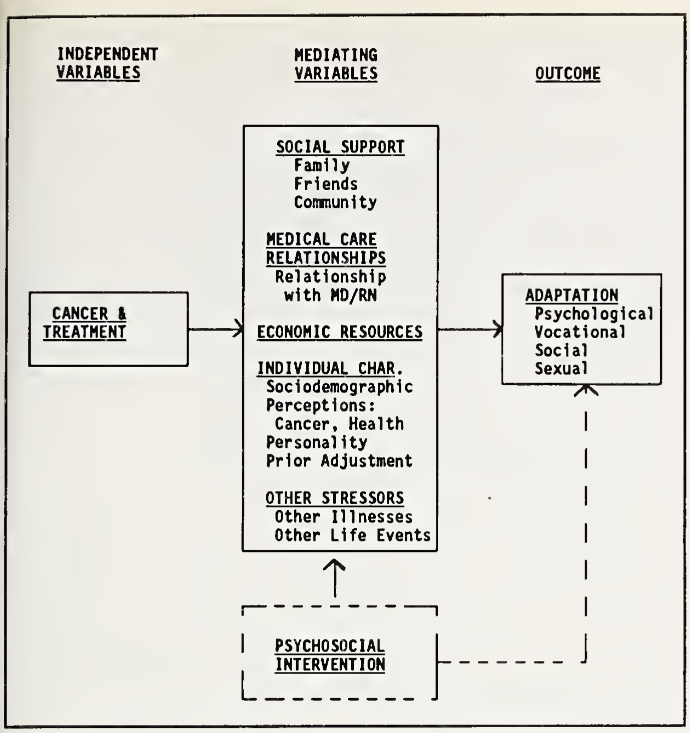

Fig. 1. Vulnerability model of patients' adaptation to cancer.

come. Mediating factors that may influence patient adaptation are included in the model, such as social support, relationship to the health care team, economic resources, personality characteristics, concurrent stressful life events, and comorbid conditions. In addition to assessing social support as a potential buffer serving to protect patients from the impact of stress, the supportive role of the medical team to a patient's adjustment, often overlooked by researchers, is highlighted in this model. Psychosocial interventions must be developed to affect these mediating variables or QOL itself.

While no single study can incorporate the measurement of all mediating variables, the vulnerability model has served to guide the selection of variables and, consequently, instruments of important mediating factors of adaptation in a number of our studies: family environment and health beliefs in long-term psychosocial adaptation of leukemia survivors (CALGB 8963) (*16*); preference for control over health care in a study of patient-controlled analgesia for patients with severe pain (CALGB 8872, in collaboration with the Cancer Control Committee) (*66*); the role of stressful life events, social support, and spirituality in the psychosocial adaptation of breast cancer patients, as described above (CALGB 9364); and the patient's relationship with his or her medical team in a study of the treatment of fever and neutropenia at home by antibiotics (CALGB 9170, in collaboration with the Cancer Control Committee). By identifying critical factors that exacerbate patients' vulnerability to the stresses of cancer, as well as those that are protective, and examining the balance of these forces within the context of this model, wide differences in patient adaptation to the same disease and treatment can be better understood.

Guidelines for Future QOL Research

Prioritizing Clinical Trials for QOL Study

Cost containment today mandates prioritizing which studies receive a QOL component. In the CALGB, outside funding must be obtained for any study that exceeds the capacity of the single, National Cancer Institute-supported research interviewer, which is approximately two studies at a time, with several assessments during and off treatment. Given the volume of interest in QOL research, one interviewer has not been adequate to meet the demand. When there have been additional studies of interest, support of approximately $25 000 per year has been sought to cover the cost of telephone-derived QOL data collection. Apportioning QOL studies by disease site or modality (i.e., one per committee) is one way to place a limit on QOL studies, but this does not take into account the possibility of several studies with important QOL issues becoming simultaneously active within a single committee. A mechanism must be established to prioritize QOL studies, involving both oncologists and psycho-oncologists within the cooperative clinical trial group, similar to the designation of high-priority clinical trials. However, prioritizing QOL studies will not eliminate the need for financial support for this research. As of today, the CALGB remains the only cooperative group with a budget dedicated to QOL research. Although minimal, that budget has been pivotal in our research effort over the years. With the prioritization of QOL studies, in conjunction with the appropriate support, resources can then be allocated so that studies are conducted with the proper methodologic rigor and depth of measurement.

Data Collection Method

We consider the telephone interview as the QOL data collection method of choice for cooperative clinical trials groups. QOL information is validly obtained via telephone interview, yields minimal missing data, results in excellent retention of patients for follow-up assessments, and has high patient satisfaction. The high rate of successfully completed interviews in the breast cancer megestrol acetate study (CALGB 8864) (*34*) was felt, in part, to be due to the rapport that developed between interviewer and patient over the course of the three assessments. Retaining patient compliance to repeated assessments in QOL studies could thus be enhanced as a consequence of the rapport between interviewer and patient. Computer-assisted telephone interviewing (CATI) (*26,67*) could be used to reduce some of the additional costs of telephone interviewing by increasing efficiency in coding and data processing, without the visible presence of computer technology interfering with the interview process. The use of mailed questionnaires may be appealing at the outset because of ease of administration and lowest cost of all the data collection methods (*27,30-32*). Self-administration of questionnaires in the clinic is similarly viewed favorably by those in the cooperative clinical trials groups for those reasons, although there are hidden costs that render this approach not so inexpensive, as Moinpour's paper in this monograph suggests. Excellent completion rates using self-administered questionnaires in the clinic have been reported by the Southwest Oncology Group (*68*) and the National Cancer Institute of Canada Clinical Trials Group (*69*). However, that certainly has not been

our experience using this method of data collection, which proved quite scientifically costly to the CALGB in its early studies, nor has it been the experience of other cooperative clinical trials groups, such as the European Organization for Research and Treatment of Cancer (*70,71*). While mixed-mode methods would certainly improve compliance by an average of 5%-15% (*21*), it is not clear if that would sufficiently offset the magnitude of the problem of missing data, particularly as patients' functioning deteriorates and clinic attendance becomes sporadic. Therefore, when the budget for a QOL study is being developed, a broader view of cost needs to be adopted that includes our confidence in obtaining results on which the scientific community can build.

Measurement

There is a current wave of conservatism and demand for simplification in QOL measurement, with the suggested use of a single measure across all clinical trials, with perhaps a few additional items specific to a protocol. This trend is due to multiple factors: 1) frustration stemming from our current inability to compare results across trials due to variability in measurement (*72*), 2) a lack of understanding as to how to evaluate different measures for their appropriateness in measuring specific QOL issues in the different trials, and 3) an increasing limitation in financial resources devoted to QOL research. Indeed, the use of a core measure will allow for comparisons across clinical trials and provide much needed information concerning QOL issues for specific patient populations through the ongoing development of a database. However, the paramount consideration when selecting QOL measures for a clinical trial is that it serve as a valid test of the QOL research question, *specific* to that trial. Despite the clear benefits of using a common measure across trials, this issue can never supersede in importance the primary scientific mandate: answering the QOL question for that trial.

As the oncology community's confidence in the availability of valid QOL measures has increased, attention has begun to shift to understanding the *clinical* significance of statistically significant results. For many QOL measures, the clinical significance of findings is not readily apparent, nor are normative data available from either large community samples or relevant cancer patient populations to provide a frame of reference for interpreting patients' scores. The clinical significance of QOL scores will be determined as normative information is obtained for these measures and correlated with clinically well-understood, disease-relevant measures, such as the Karnofsky performance status scale, psychiatric diagnoses, and behavioral indicators of psychosocial functioning (*65*). The cooperative clinical trials groups are the ideal context within which to conduct this research, given the broad representation of patient populations and documentation of disease and treatment variables.

Cost Analysis

The cost-conscious atmosphere in health care in the past 5 years has resulted in an increased interest in including cost analyses in relation to survival, toxicity, and/or QOL in the evaluation of cancer treatments (*73-76*). However, as of yet, it is rare to see all four end points included in the research design. The CALGB's newly created Clinical Economics Working Group, in collaboration with the Psycho-Oncology Committee, will select studies in which there are clear cost as well as QOL implications for different treatments. Our first effort in this area will be a study of the hepatic arterial infusion pump, in which colorectal cancer patients with hepatic metastases will be randomly assigned to receive chemotherapy either by a surgically implanted pump or systemic therapy (CALGB 9481). The significance of this model approach to treatment evaluation is that four major parameters are included: survival, toxicity, QOL, and cost, creating an enriched dataset from which to understand the impact of a cancer treatment on patients' lives.

Conceptual Issues Concerning QOL Research

QOL research has not been theory driven, but rather, it has been guided by hypotheses as to which treatment arm will result in worse QOL, based on expected side effects, treatment efficacy, or other treatment-related effects. This has been appropriate, given the nature of the research questions and the measurement limitations imposed by patients' level of illness. However, by having a paucity of theoretical issues tested within the trials, little light has been thrown on identifying specific psychosocial mechanisms by which cancer patients adjust to their disease and treatment. Because planning rational interventions will require such information, theoretical models need to be considered in the development of QOL research in clinical trials. By concentrating resources in identified high-priority studies, expanded measurement is made more possible, enabling the testing of theoretically based questions.

Conclusion

The most significant clinical application of QOL research in phase III clinical trials will be to assist both patients and their oncologists in making treatment decisions by providing them with relevant information concerning a specific treatment's impact on QOL. QOL issues are routinely taken into account in decision-making: oncologists, on the basis of their clinical experience, and patients, on the basis of their judgment about potential efficacy versus expected side effects and disruption in function. When QOL research is conducted with the proper thought and methodologic rigor, the combined effect of obtaining QOL with survival, toxicity, and cost data in the evaluation of cancer treatments in clinical trials will be to guide treatment decisions from a more rational perspective.

References

(*1*) Holland JC, Silberfarb P, Tross S, Cella D. Psychosocial research in cancer: the Cancer and Leukemia Group B [CALGB] experience. In: Ventafridda V, van Dam FS, Yancik R, Tamburini M, editors. Assessment of quality of life and cancer treatment. Amsterdam: Excerpta Medica 1986:89-101.

(*2*) Ahles TA, Silberfarb PM, Rundle AC, Holland JC, Kornblith AB, Canellos GP, et al. Quality of life in patients with limited small-cell carcinoma of the lung receiving chemotherapy with or without radiation therapy, for the Cancer and Leukemia Group B. Psychother Psychosom 1994;62:193-9.

(*3*) Silberfarb PM, Holland JC, Anbar D, Bahna G, Maurer LH, Chahinian AP, et al. Psychological response of patients receiving two drug regimens for lung carcinoma. Am J Psychiatry 1983;140:110-1.

(4) Cella DF, Orav EJ, Kornblith AB, Holland JC, Silberfarb PM, Lee KW, et al. Socioeconomic status and cancer survival. J Clin Oncol 1991;9:1500-9.
(5) Holland JC, Korzun AH, Tross S, Cella DF, Norton L, Wood W. Psychosocial factors and disease free survival (DFS) in stage II breast carcinoma. Proc ASCO 1986;5:237.
(6) Silberfarb PM, Anderson KM, Rundle AC, Holland JC, Cooper MR, McIntyre OR. Mood and clinical status in patients with multiple myeloma. J Clin Oncol 1991;9:2219-24.
(7) Holland JC, Korzun AH, Tross S, Silberfarb P, Perry M, Comis R, et al. Comparative psychological disturbance in patients with pancreatic and gastric cancer. Am J Psychiatry 1986;143:982-6.
(8) Cella DF, Jacobsen PB, Orav EJ, Holland JC, Silberfarb PM, Rafla S. A Brief POMS measure of distress for cancer patients. J Chronic Dis 1987;40:939-42.
(9) Cella DF, Tross S, Orav EJ, Holland JC, Silberfarb PM, Rafla S. Mood states of patients after the diagnosis of cancer. J Psychosocial Oncol 1989; 7(1/2):45-54.
(10) Cella DF, Orofiamma B, Holland JC, Silberfarb PM, Tross S, Feldstein M, et al. The relationship of psychological distress, extent of disease, and performance status in patients with lung cancer. Cancer 1987;60:1661-7.
(11) Rowland JH, Glidewell OJ, Sibley RF, Holland JC, Tull R, Berman A, et al. Effects of different forms of central nervous system prophylaxis on neuropsychological function in childhood leukemia. J Clin Oncol 1984;2: 1327-35.
(12) Kornblith AB, Anderson J, Cella DF, Tross S, Zuckerman E, Cherin E, et al. Quality of life assessment of Hodgkin's disease survivors: a model for cooperative clinical trials. Oncology 1990;4:93-101.
(13) Kornblith AB, Anderson J, Cella DF, Tross S, Zuckerman E, Cherin E, et al. Hodgkin's disease survivors at increased risk for problems in psychosocial adaptation. The Cancer and Leukemia Group B. Cancer 1992;70:2214-24.
(14) Kornblith AB, Anderson J, Cella DF, Tross S, Zuckerman E, Cherin E, et al. Comparison of psychosocial adaptation and sexual function of survivors of advanced Hodgkin's disease treated by MOPP, ABVD, or MOPP alternating with ABVD. Cancer 1992;70:2508-16.
(15) Janov AJ, Anderson J, Cella DF, Zuckerman E, Kornblith AB, Holland JC, et al. Pregnancy outcome in survivors of advanced Hodgkin's disease. Cancer 1992;70:688-92.
(16) Greenberg DB, Herndon JE, Kornblith AB, Zuckerman E, Schiffer CA, Weiss RB, et al. Long-term psychosocial adaptation of survivors of acute leukemia. Proc ASCO 1995;14:508.
(17) Johnson JR, Temple R. Food and Drug Administration requirements for approval of new anticancer drugs. Cancer Treat Rep 1985;69:1155-7.
(18) Kornblith AB, Holland JC. Handbook of measures for psychological, social and physical function in cancer. Volume I: Quality of life. New York (NY): Memorial Sloan-Kettering Cancer Center, 1994.
(19) Lavrakas PJ. Telephone survey methods: sampling, selection and supervision. Newbury Park (CA): Sage, 1987.
(20) Herzog AR, Rodgers WL, Kulka RA. Interviewing older adults: a comparison of telephone and face-to-face modalities. Public Opin Q 1983; 47:405-18.
(21) Herzog AR, Rodgers WL. The use of survey methods in research on older Americans. In: Wallace RB, Woolson RF, editors. The Epidemiologic Study of the Elderly. New York: Oxford Univ Press, 1992:60-90.
(22) Dillman DA. Mail and telephone surveys: the total design method. New York: John Wiley & Sons, 1978.
(23) Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA. Telephone versus in-person surveys of community health status. Am J Public Health 1982;72:1017-21.
(24) Aneshensel CS, Yokopenic PA. Tests for the comparability of a causal model of depression under two conditions of interviewing. J Pers Soc Psychology 1985;49:1337-48.
(25) Colombotos J. Personal versus telephone interviews: effect on responses. Public Health Rep 1969;84:773-82.
(26) Groves R, Kahn M. Surveys by telephone: a national comparison with personal interviews. New York: Academic Press, 1979.
(27) Hochstim JR. A critical comparison of three strategies of collecting data from households. J Am Stat Assoc 1967;62:976-89.
(28) Jordan LA, Marcus AC, Reeder LG. Response styles in telephone and household interviewing: a field experiment. Public Opin Q 1980;44:210-22.
(29) Locander W, Sudman S, Bradburn N. An investigation of interview method, threat and response distortion. J Am Stat Assoc 1976;71:269-74.
(30) Siemiatycki J. A comparison of mail, telephone, and home interview strategies for household health surveys. Am J Public Health 1979;69:238-45.
(31) Warner JL, Berman JJ, Weyant JM, Ciarlo JA. Assessing mental health program effectiveness: a comparison of three client follow-up methods. Eval Rev 1983;7:635-58.
(32) Walker AH, Restuccia JD. Obtaining information on patient satisfaction with hospital care: mail versus telephone. Health Serv Res 1984;19:83-98.
(33) Fobair P, Hoppe RT, Bloom J, Cox R, Varghese A, Spiegel D. Psychosocial problems among survivors of Hodgkin's disease. J Clin Oncol 1986;4: 805-14.
(34) Kornblith AB, Hollis DR, Zuckerman E, Lyss AP, Canellos GP, Cooper MR, et al. Effect of megestrol acetate on quality of life in a dose-response trial in women with advanced breast cancer. J Clin Oncol 1993;11:2081-9.
(35) Holland JC, Kornblith AB, Zuckerman E. A centralized model for quality of life (QOL) data collection by telephone interview in multicenter clinical trials. Proc ASCO 1992;11:406.
(36) Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality-of-life assessment in cancer clinical trials. Recent Results Cancer Res 1988;111:231-49.
(37) Schipper H, Clinch J, McMurray A, Levitt M. Measuring the quality of life of cancer patients: the Functional Living Index-Cancer: development and validation. J Clin Oncol 1984;2:472-83.
(38) Kosty MP, Fleishman SB, Herndon JE 2d, Coughlin K, Kornblith AB, Scalzo A, et al. Cisplatin, vinblastine, and hydrazine sulfate in advanced, non-small-cell lung cancer: a randomized placebo-controlled, double-blind phase III study of the Cancer and Leukemia Group B. J Clin Oncol 1994;12:1113-20.
(39) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.
(40) Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. Eur J Cancer 1994;30A:635-42.
(41) Broadhead WE, Gehlbach SH, de Gruy FV, Kaplan BH. The Duke-UNC Functional Social Support Questionnaire. Measurement of social support in family medicine patients. Med Care 1988;26:709-23.
(42) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy Scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.
(43) Brook RH, Ware JE Jr, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, et al. Conceptualization and measurement of health for adults in the Health Insurance Study: overview (Vol. VIII). Santa Monica (CA): Rand Corporation, October 1979.
(44) Veit CT, Ware JE Jr. The structure of psychological distress and well-being in general populations. J Consult Clin Psychol 1983;51:730-42.
(45) McCorkle R, Young K. Development of a symptom distress scale. Cancer Nurs 1978;1:373-8.
(46) McCorkle R, Quint-Benoliel J. Symptom distress, current concerns and mood disturbance after diagnosis of life-threatening disease. Soc Sci Med 1983;17:431-8.
(47) Portenoy RK, Thaler HT, Kornblith AB, Lepore JM, Friedlander-Klar H, Kiyasu E, et al. The Memorial Symptom Assessment Scale: an instrument for the evaluation of symptom prevalence, characteristics and distress. Eur J Cancer 1994;30A:1326-36.
(48) Cleeland CS, Syrjala KL. How to assess cancer pain. In: Turk D, Melzack R, editors. Pain assessment. New York: Guilford Press, 1992:360-87.
(49) Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain 1983;17:197-210.
(50) Derogatis LR, Melisaratos N. The DFSI: a multidimensional measure of sexual functioning. J Sex Marital Ther 1979;5:244-81.
(51) Derogatis LR. The Psychosocial Adjustment to Illness Scale (PAIS). J Psychosom Res 1986;30:77-91.
(52) Derogatis LR, Melisaratos N. The Brief Symptom Inventory: an introductory report. Psychol Med 1983;13:595-605.
(53) Horowitz M, Wilner N, Alvarez W. Impact of Event Scale: a measure of subjective stress. Psychosom Med 1979;41:209-18.
(54) Becker MH, Maiman LA. Sociobehavioral determinants of compliance with health and medical care recommendations. Med Care 1975;18(1):10-24.
(55) Kirscht JP. The Health Belief Model and predictions of health actions. In: Gochman DS, editor. Health behavior: emerging research perspectives. New York: Plenum Press, 1988:27-41.
(56) Leventhal H, Diefenbach M, Leventhal EA. Illness cognition: used common sense to understand treatment adherence and affect cognition interactions. Cognitive Ther Res 1992;16:143-63.
(57) Weinstein ND. Testing four competing theories of health-protective behavior. Health Psychol 1993;12:324-33.
(58) Prochaska JO, DiClemente CC. The transtheoretical approach: crossing traditional boundaries of change. Homewood (IL): Dorsey Press, 1984.
(59) Sherbourne CD, Stewart AL. The MOS Social Support Survey. Soc Sci Med 1991;32:705-14.

(*60*) Halper MS, Winawer SJ, Brody RS, Andrews M, Roth D, Burton G. Issues of patient compliance. In: Winawer SJ, Schottenfeld DM, Sherlock P, editors. Colorectal cancer: prevention, epidemiology, and screening. New York: Raven Press, 1980:299-310.
(*61*) Fox BH. The role of psychological factors in cancer incidence and prognosis. Oncology 1995;9:245-53.
(*62*) Sarason IG, Johnson JH, Siegel JM. Assessing the impact of life changes: development of the Life Experiences Survey. J Consult Clin Psychol 1978;46:932-46.
(*63*) Fashingbauer TR, Zisook S, DeVaul R. The Texas Revised Inventory of Grief. In: Zisook S, editor. Biopsychosocial Aspects of Bereavement. Washington, DC: American Psychiatric Press, 1987:111-24.
(*64*) Dohrenwend BS, Dohrenwend BP. Life stress and illness: formulation of the issues. In: Dohrenwend BS, Dohrenwend BP, editors. Stressful life events and their contexts. New York: Prodist, 1981:1-27.
(*65*) Kornblith AB, Thaler HT, Wong G, Vlamis V, McCarthy-Lepore J, Loseth DB, et al. Quality of life of women with ovarian cancer. Gyn Oncol. In press.
(*66*) Citron M, Conaway M, Zhukovsky D, Kornblith AB, Berkowitz I, Pascall V, et al. Efficacy of patient-controlled analgesia (PCA) vs. continuous intravenous morphine (CIVM) for the treatment of severe cancer pain: CALGB 8872. Proc ASCO 1993;12:433.
(*67*) Patrick DL, Erickson P. Health status and health policy: allocating resources to health care. New York: Oxford Univ Press, 1993.
(*68*) Hayden KA, Moinpour CM, Metch B, Feigl P, O'Bryan RM, Green S, et al. Pitfalls in quality-of-life assessment: lessons from a Southwest Oncology Group breast cancer clinical trial. Oncol Nurs Forum 1993;20:1415-9.
(*69*) Sadura A, Pater J, Osoba D, Levine M, Palmer M, Bennett K. Quality-of-life assessment: patient compliance with questionnaire completion. J Natl Cancer Inst 1992;84:1023-6.
(*70*) da Silva FC. Quality of life in prostatic cancer patients. Cancer 1993; 72:3803-6.
(*71*) Aaronson NK. Quality of life research in cancer clinical trials: a need for common rules and language. Oncology 1990;4(5):59-66.
(*72*) Gill TM, Feinstein AR. A critical appraisal of the quality-of-life measurements. JAMA 1994;272:619-26.
(*73*) Feeny D, Labelle R, Torrance GW. Integrating economic evaluations and quality of life assessments. In: Spilker B, editor. Quality of life assessments in clinical trials. New York: Raven Press, 1990:71-83.
(*74*) Goodwin PJ, Feld R, Evans WK, Pater J. Cost-effectiveness of cancer chemotherapy: an economic evaluation of a randomized trial in small-cell lung cancer. J Clin Oncol 1988;10:1537-47.
(*75*) Bennett CL, Armitage JL, Armitage GO, Vose JM, Bierman PJ, Armitage JO, et al. Costs of care and outcomes for high-dose therapy and autologous transplantation for lymphold malignancies: results from the University of Nebraska 1987 through 1991. J Clin Oncol 1995;13:969-73.
(*76*) Weeks JC, Tierney MR, Weinstein MC. Cost-effectiveness of prophylactic intravenous immune globulin in chronic lymphocytic leukemia. N Engl J Med 1991;325:81-6.

Note

Supported by Public Health Service grant CA31946 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services; the T. J. Martell Foundation; the John D. and Catherine T. MacArthur Foundation; the Bristol-Myers Pharmaceutical Company; and the Chemotherapy Foundation.

A Cooperative Group Report on Quality-of-Life Research: Lessons Learned

*Mary S. McCabe**

Introduction

This paper is a summary of the presentations given by representatives of the National Cancer Institute Cooperative Groups at the March 1-2, 1995, meeting, "Workshop on Quality of Life in Clinical Cancer Trials." The individual sections convey the diverse interests, unique patient populations, modality focus, and overall thoughtful approaches that each of the cooperative groups brings to oncology quality-of-life (QOL) research (Table 1).

This summary provides a unique opportunity to review the problems, successes, and plans for the conduct of QOL research from a national perspective. The study-specific discussions are intended to provide clinically valuable information and to assist in the planning of future QOL evaluations that will advance the field and ultimately answer questions that will benefit patients with cancer.

**Correspondence to:* Mary S. McCabe, R.N., National Institutes of Health, EPN, Rm. 715, Bethesda, MD 20892.

Table 1. Active cooperative group treatment trials with QOL end points*,†

Protocol/ coordination center No.	Title	Phase	QOL instrument
	Disease group—AIDS		
E1493	Sequential chemotherapy and radiotherapy for AIDS-related primary central nervous system lymphoma	II	Functional Assessment of HIV Infection, HIV QOL Survey, HIV Questionnaire
	Disease group—breast		
CALGB-9342	Study of Taxol (paclitaxel) at three dose levels in the treatment of patients with metastatic breast cancer	III	Functional Living Index—Cancer, Symptom Distress Scale
E1193	Trial of doxorubicin versus paclitaxel versus paclitaxel plus adriamycin plus granulocyte-colony stimulating factor in metastatic breast cancer	III	Functional Assessment of Cancer—Breast Cancer
INT-0121/EST-2190	Study of conventional adjuvant chemotherapy versus high-dose chemotherapy and autologous bone marrow transplant as questionnaire/adjuvant intensification therapy following conventional adjuvant chemotherapy in patients with stage II and III breast cancer at high risk of recurrence	III	Breast Chemotherapy
INT-0142/E-3193	Comparison of tamoxifen versus tamoxifen with ovarian ablation in premenopausal women with axillary node-negative receptor-positive breast cancer <2 cm	III	Functional Assessment of Cancer Therapy—Breast, ECOG Menopausal Symptom Form
T90-0180/PBT-1	Rrandomized comparison of maintenance chemotherapy with CTX, MTX, and 5-FU versus high-dose chemotherapy with CTX, thiotepa, and CBDCA and ABMT support for women with metastatic breast cancer responding to conventional induction chemotherapy	III	Medical Outcomes Study Short Form-20, Symptom Distress Scale, Profile of Mood States, Mental Adjustment to Cancer Scale
	Disease group—central nervous system		
INT-0140/POG-9331	Treatment of children with early-stage medulloblastoma: standard-dose craniospinal irradiation versus reduced-dose craniospinal irradiation plus adjuvant chemotherapy with cisplatin, cyclophosphamide, and vincristine	III	POG QOL Questionnaire
INT-0149/RTOG-9402	Intergroup randomized comparison of radiation alone versus pre-radiation chemotherapy for pure and mixed anaplastic oligodendrogliomas	III	Karnofsky Performance Scale, Mini-Mental State Examination, EORTC-B QOL Questionnaire
NCCTG-93-72-52	Trial of BCNU and cisplatin versus BCNU alone and standard radiation therapy versus accelerated radiation therapy in patients with high-grade glioma	III	Mini-Mental State Examination, Neurologic Function Status
RTOG-93-05	Trial comparing the use of radiosurgery followed by conventional radiotherapy with BCNU to conventional radiotherapy with BCNU for supratentorial glioblastoma multiforme	III	Mini-Mental State Examination, Spitzer QOL Index
RTOG-94-11	Ttumor volume-influenced dose escalation of accelerated hyper-fractionated radiotherapy to 64.0 and 70.4 Gy with BCNU for newly diagnosed radiosurgery-ineligible glioblastoma multiforme patients	II	Mini-Mental State Examination

Table 1 (continued). Active cooperative group treatment trials with QOL end points*,†

Protocol/ coordination center No.	Title	Phase	QOL instrument
RTOG-94-17	Single-arm, open label, study of intravenously administered tirapazamine plus radiation therapy for high-grade glioblastoma multiforme	II	Mini-Mental State Examination
	Disease group—gastrointestinal		
INT-0146/NCCTG-93-46-53	Prospective randomized trial comparing laparoscopic-assisted colectomy versus open colectomy for colon cancer	III	Symptom Distress Scale, QOL-Index, Q-TWiST
INT-0147/RTOG-9401	Intergroup randomized trial of preoperative versus postoperative combined modality therapy for resectable rectal cancer	III	Anorectal Function Assessment Tool, Functional Assessment of Therapy—Cancer
NCCTG-91-46-51	Salvage protocol for patients with advanced colorectal cancer who have relapsed following surgical adjuvant chemotherapy	II	QOL Uniscale
	Disease group—genitourinary		
CALGB-9182	Randomized comparison of low-dose steroids and mitoxantrone versus low-dose steroids in patients with "hormone refractory" stage D2 carcinoma of the prostate	III	Functional Living Index—Cancer, Symptom Distress Scale, Sexual and Urologic Functioning Questionnaire, Problems in Daily Activities
E7892	Randomized, double-blinded trial of adjuvant hormonal therapy for surgically treated pathologic stage C carcinoma of the prostate	III	Functional Assessment of Cancer Therapy—Prostate
INT-0162/SWOG-9346	Intermittent androgen deprivation in patients with stage D2 prostate cancer	III	Medical Outcomes Study Short Form-36, Medical Outcomes Study Short Form-20, Symptom Distress Scale, Linear Analogue Self Assessment
RTOG-94-08	Phase III trial of the study of endocrine therapy used as a cytoreductive and cytostatic agent prior to radiation therapy in good prognosis, locally confined adenocarcinoma of the prostate	III	Sexual Adjustment Questionnaire
T94-0110/NCIC-CTG	Intergroup (NCIC CTG, and ECOG) phase III randomized trial comparing total androgen blockade versus total androgen blockade plus pelvic irradiation in clinical stage T3-4, N0, M0 adenocarcinoma of the prostate	III	EORTC Core Questionnaire-33, Functional Assessment of Cancer Therapy—Prostate
	Disease group—gynecologic		
E2E93	Clinical trial of an outpatient paclitaxel and carboplatin regimen in the treatment of suboptimally debulked epithelial carcinoma of the ovary	II	Functional Assessment of Cancer Therapy—Ovarian
GOG-145	Randomized study of surgery versus surgery plus vulvar radiation in the management of poor prognosis primary vulvar cancer and of radiation versus radiation and chemotherapy for positive inguinal nodes	III	Functional Assessment of Cancer Therapy—General, GOG Symptom Inventory, Groningen Arousability Scale, Groningen Body Image Scale
GOG-152	Randomized study of cisplatin (NSC 119875) and paclitaxel (NSC 125973) with interval secondary cytoreduction versus cisplatin and paclitaxel in patients with suboptimal stage III and IV epithelial ovarian carcinoma	III	Functional Assessment of Cancer Therapy—Ovarian
GOG-9102	Effect of alopecia on cancer patients' body image and the role of audiovisual information on body image	Other	Secourd and Jourard Body Cathexis Index
GOG-LAP1	Orientation and evaluation study of surgeon proficiency in performing a GOG standardized procedure for laparoscopic FIGO staging in adenocarcinoma of the endometrium	Other	Functional Assessment of Cancer Therapy—General, Medical Outcomes Survey—Physical Functioning Subscale, Wisconsin Brief Pain Inventory, Fear of relapse/Recurrence Scale, Sexual Functioning Scale, Body Image
SWOG-9324	Trial of vinorelbine tartrate (navelbine) for patients with relapsed ovarian cancer	II	Medical Outcomes Study Short Form-36
	Disease group—head and neck		
RTOG-90-03	Randomized study to compare twice daily hyperfractionation, accelerated hyperfractionation with a split and accelerated fractionation with concomitant boost to standard fractionation radiotherapy for squamous cell carcinomas of head and neck	III	Functional Assessment of Cancer Therapy—Head and Neck, List Performance Status Scale, Dische Morbidity Scoring Tool
RTOG-91-11	Trial to preserve the larynx: induction chemotherapy and radiation therapy versus concomitant chemotherapy and radiation therapy versus radiation therapy	III	Functional Assessment of Cancer—Head and Neck, Symptom Scale

Table 1 (continued). Active cooperative group treatment trials with QOL end points*,†

Protocol/ coordination center No.	Title	Phase	QOL instrument
	Disease group—leukemia		
CCG-1941	Bone marrow transplantation versus prolonged intensive chemotherapy for children with acute lymphoblastic leukemia after an initial bone marrow relapse	III	Ontario Health Survey
POG-9421	Evaluation of standard versus high-dose ARA-C induction followed by the randomized use of cyclosporine A as an MDR reversal agent, compared with allogeneic BMT, in childhood AML	III	Bayley Scales of Infant Development, Vineland Adaptive Behavior Scales, POG QOL Battery, Family Environment Scale, Wechsler Intelligence Scale for Children-III, Beery Test of Visual-Motor Integration, Achenbach CBC, Wide Range Achievement Test
	Disease group—lung		
E3592	Cisplatin plus etoposide versus daily oral etoposide in elderly patients with extensive-stage small cell lung cancer	III	Functional Assessment of Cancer Therapy—Lung, ECOG Neurotoxicity—Related QOL Questionnaire
E4593	Study of hyperfractionated accelerated radiation therapy for advanced, unresectable non-small-cell lung cancer with or without G-CSF	II	Functional Assessment of Cancer Therapy—Lung
E7593	Cisplatin plus etoposide versus cisplatin plus etoposide followed by topotecan in extensive-stage small cell lung cancer	III	Funcational Assessment of Cancer Therapy—Lung
INT-0131	Randomized study of CODE plus thoracic irradiation versus alternating CAV and EP for extensive stage small-cell lung cancer	III	EORTC Quality of Life Questionnaire
NCCTG-89-20-51	Study in extensive-disease small-cell lung cancer to evaluate the addition of megestrol acetate to the etoposide/cisplatin regimen	III	Functional Living Index—Cancer
	Disease group—lymphoma		
SWOG-9133	Randomized trial of subtotal nodal irradiation versus doxorubicin, vinblastine and subtotal nodal irradiation for stage I-IIA Hodgkin's disease	III	CARES-SF, Symptom Distress Scale, Medical Outcomes Study Short Form-36
SWOG-9208	Health status and QOL in patients with early stage Hodgkin's disease	Other	Symptom Distress Scale, Cancer Rehabilitation Evaluation System—Short Form, Medical Outcomes Study Short Form-36
	Disease group—multiple sites		
INT-0143/RTOG-9310	Intergroup phase II combined modality treatment of primary central nervous system lymphoma	II	Mini-Mental State Examination
	Disease group—myelodysplastic syndrome		
CALGB-9221	Randomized phase III controlled trial of subcutaneous 5-azacytidine (NSC #102816) versus observation in myelodysplastic syndromes	III	EORTC QOL Questionnaire, Revised Rand General Well-Being Scale

*Source: Cancer Therapy Evaluation Program database of cooperative group trials.
†CTX = cyclophosphamide; MTX = methotrexate; 5-FU = fluorouracil; CBDCA = carboplatin; BCNU = carmustine; NCIC CTG = National Institute of Canada-Clinical Trials Group; ECOG = Eastern Cooperative Oncology Group; GOG = Gynecologic Oncology Group; ARA-C = cytarabine; MDR = multidrug resistance; BMT = bone marrow transplant; AML = acute myeloid leukemia; CAV = cyclophosphamide, doxorubicin, and vincristine; and EP = etoposide and cisplatin.

Cancer and Leukemia Group B (CALGB)

Alice B. Kornblith

Overview

The past 5 years have been very active for the Psycho-Oncology Committee. Studies were conducted by or are currently active with five Disease Committees: 1) breast (CALGB-8082, CALGB-8864, CALGB-9066, CALGB-9342, and CALGB-9364), 2) lung (CALGB-8931 and CALGB-9033), 3) lymphoma (CALGB-8561, CALGB-8562, and CALGB-9497), 4) leukemia (CALGB-8963 and CALGB-9221), and 5) prostate (CALGB-9181 and CALGB-9182). Three studies are currently in development with the Gastrointestinal Committee. Three major areas of research have been pursued since 1990: 1) quality of life (QOL) of patients on active treatment (CALGB-8083, CALGB-8534, CALGB-8864, CALGB-8872, CALGB-8931, CALGB-8971, CALGB-9033, CALGB-9066, CALGB-9181, CALGB-9182, CALGB-9221, and CALGB-9342), 2) psychosocial adaptation of leukemia and Hodgkin's disease survivors (CALGB-8963, CALGB-8561, CALGB-8562, and CALGB-9497), and 3) psychosocial and sociodemographic factors as predictors of survival (CALGB-7761, CALGB-8082, and CALGB-9364). Eight journal articles (*1-8*) and six abstracts (*9-13,16*) have been published. New initiatives for the Psycho-Oncology Committee address the development of interventions: 1) telephone counseling to improve patients' adjustment during active treatment or upon completing treatment, 2) improving doctor–patient communication, and 3) use of patient advocates to improve minority participation in clinical trials. Furthermore, we will be collaborating with the newly created Clinical Economics Working Group in selected studies in which there are clear cost as well as QOL implications of different treatments (e.g., hepatic arterial infusion protocol CALGB-9481). Last, with the increasing development of methods for genetic testing and other cancer-screening methods, QOL issues concerning individuals at high risk for cancer have assumed critical importance. The study of the psychological consequences of intensified screening of relatives at high risk for colon cancer will serve as a paradigm for this research area (9X6Q). Additional QOL research related to patients' participation in genetic research is currently being explored across the Psycho-Oncology, Cancer Control, and Oncology Nursing Committees.

QOL During Active Treatment

Active Protocols

CALGB-9182—Randomized comparison of low-dose steroids and mitoxantrone versus low-dose steroids in patients with hormone refractory stage D2 carcinoma of the prostate. This study uses telephone interviewing as the method for data collection. The QOL measures in CALGB-9181 and CALGB-9182 are identical. QOL is being evaluated in both protocols by use of the Functional Living Index: Cancer McCorkle Symptom Distress Scale, sexual and urological functioning subscales of the EORTC (i.e., European Organization for Research and Treatment of Cancer) Prostate Questionnaire, Rand Functional Limitations Scale (modified), and the interference of pain with daily functioning subscale of the Wisconsin Brief Pain Inventory.

CALGB-9221—A randomized phase III controlled trial: subcutaneous 5-azacytidine versus observation in myelodysplastic syndromes. The QOL hypothesis in this study is that those randomly assigned to the 5-azacytidine arm will experience a better QOL because of better symptom control (e.g., fewer hospitalizations, fewer infections, and decreased fatigue) than those in the control group.

CALGB-9334—Sclerosis of pleural effusions by talc thoracoscopy versus talc slurry: a phase III study. This study will compare the QOL of patients with pleural effusions randomly assigned to receive talc slurry, administered at the bedside, or talc thoracoscopy, conducted in the operating room. Patients' QOL will be assessed using the EORTC QLQ-C30 Questionnaire, and items will be developed to assess patients' satisfaction with these two procedures. In addition, pain will be assessed daily with the use of a visual analogue scale, until the chest tube is removed.

CALGB-9342—Phase III study of paclitaxel (Taxol) at three dose levels in the treatment of patients with metastatic breast cancer. The objectives of the QOL component of this study are to examine the prognostic value of QOL scores at base line and to compare patients' QOL on 175, 210, or 250 mg/m^2 paclitaxel. QOL is assessed by the Functional Living Index—Cancer (FLIC) and the McCorkle Symptom Distress Scale.

CALGB-9497—Health status and QOL in patients with early stage Hodgkin's disease: a companion study to CALGB-9391/SWOG (i.e., Southwest Oncology Group) 9133. This intergroup trial under SWOG evaluates the QOL of patients with early stage Hodgkin's disease over a 7-year period. Measures include the CARES-SF (Cancer Rehabilitation Evaluation System-Short Form), McCorkle's Symptom Distress Scale, and the MOS SF-36 (Medical Outcome Study-36 Item Short Form Health Survey) Vitality and Health Perception Subscales.

In Collaboration With Cancer Control Committee

CALGB-9170—A multicenter trial of hospital versus early discharge therapy of low-risk patients with fever and neutropenia: a phase III study. This study will compare the QOL of patients with fever and neutropenia randomly assigned to continued hospitalization or early hospital discharge with

continued care at home with antibiotics. The research nurse assesses patients' psychological state, satisfaction with medical care, and overall QOL at base line and at the end of treatment (generally within 5-7 days).

CALGB-9490—Does an oral analgesic protocol improve pain control for patients with cancer? In this intergroup trial under Eastern Cooperative Oncology Group (E4Z93/CALGB-9490), the efficacy of prescribing analgesic management of pain by protocol will be tested as a mechanism for improving pain control and the QOL of patients with metastatic or recurrent non–small-cell lung cancer, breast cancer, or multiple myeloma. Institutional sites will be randomly assigned to either the pain protocol or a standard care condition. Patients' pain, other physical symptoms, and emotional state will be assessed at base line and days 15 and 29, by use of the Brief Pain Inventory, Profile of Mood States (POMS), and the McCorkle Symptom Distress Scale.

Closed Protocols

CALGB-8083—Localized small-cell carcinoma of the lung: simultaneous chemotherapy and radiotherapy, chemotherapy versus sequential therapy (chemotherapy, radiotherapy, chemotherapy) versus chemotherapy alone. QOL and neuropsychological function of 57 patients with small-cell lung cancer were evaluated using the Trail-Making B Test (global indicator of cognitive impairment), POMS, and the Handicap Rating Scale, a physician-rated measure of five dimensions of psychosocial functioning. Patients receiving chemotherapy plus radiation therapy to both lung and brain had a significantly worse emotional state (POMS total score) and Handicap Rating Scale score at the beginning of cycle 4 of chemotherapy than those receiving chemotherapy plus prophylactic radiation therapy to the brain alone ($P<.05$). No significant differences in neuropsychological functioning were found between treatment arms (*1*).

CALGB-8534—Combination chemotherapy with intensive ACE/PCE (doxorubicin, cyclophosphamide, and etoposide/cisplatin, cyclophosphamide, and etoposide) and radiation therapy to the primary tumor and prophylactic whole-brain radiation therapy with or without warfarin in limited small-cell carcinoma of the lung: phase III. This study accrued 369 patients, and follow-up data collection has been completed. No significant differences in psychological status and neuropsychological functioning (as measured by the POMS and Trail-Making B Test) were found by the end of cycle 5 (after radiotherapy) between the two treatment arms, indicating that warfarin had no significant effect on patients' QOL.

CALGB-8864—Assessing QOL during a dose–response trial of megestrol acetate in patients with advanced breast cancer (companion to CALGB-8741). The QOL of patients with advanced breast cancer randomly assigned to receive three different doses of megestrol acetate was examined over a 3-month period. At 3 months, women treated with the lowest dose (160 mg/day) reported significantly less severe side effects ($P<.0005$) (including appetite increase, weight gain, fatigue, and feeling bloated), better physical functioning ($P<.0005$), and less psychological distress ($P = .008$) from study entry than those treated on the highest dose (1600 mg/day). No differences in body image were found among the three dose groups (*6*).

CALGB-8931—Cisplatin, vinblastine, and hydrazine sulfate (NSC-150014) in treatment of advanced non–small-cell lung cancer: a randomized, placebo-controlled, double-blinded phase III study. Patients' QOL was assessed while they were receiving either hydrazine sulfate or placebo, at 2-month intervals as long as they remained on study. Measures included the EORTC Quality of Life Questionnaire and the Duke-University of North Carolina Functional Social Support Questionnaire. At 2 months, patients receiving hydrazine sulfate had significantly worse physical symptoms and physical functioning than those receiving placebo; there were no other quality differences between the two arms (*7*).

CALGB-9033—Oral versus intravenous etoposide in combination with intravenous cisplatin in extensive small-cell lung cancer. The question of interest in this study was whether oral administration of chemotherapy improved the QOL of patients with extensive small-cell lung cancer compared with intravenous administration as a result of the ease of administration and potentially fewer side effects. Patients' QOL was assessed by telephone interview. No significant differences in QOL were found between the two treatment arms over the 3-month time period, as measured by the EORTC Quality of Life Questionnaire, the MOS Social Support Scale, and the CES-D (Center for Epidemiologic Studies-Depression Scale) Depression Scale.

CALGB-9066—QOL and psychosocial adjustment of patients with stage II or III breast cancer randomly assigned to receive high-dose CPA/cDDP (cyclophosphamide/cisplatin)/carmustine with autologous bone marrow support versus standard-dose CPA/cDDP/carmustine as consolidation to adjuvant CAF (cyclophosphamide, doxorubicin, and fluorouracil) (companion to CALGB-9082). The study's primary objective was to assess the QOL and psychosocial adaptation of stage II or III breast cancer patients randomly assigned to receive either autologous bone marrow transplant or conventional chemotherapy. Patients were interviewed by telephone using a battery of measures: the PAIS (Psychosocial Adjustment to Illness Scale), FLIC, and McCorkle Symptom Distress Scale. Follow-up data collection will continue for the next 3 years.

CALGB-9181—Randomized phase II study comparing standard-dose with moderately high-dose megestrol acetate in patients with advanced prostate cancer. The methodology used in CALGB-9181 involving telephone interviewing as the method for data collection and all QOL measures were identical to those used for CALGB-9182 (*see* CALGB-9182 above). Follow-up data collection is continuing.

In Collaboration With the Cancer Control Committee

CALGB-8872—Randomized study of patient-controlled analgesia versus continuous intravenous morphine for severe pain. The study's objective was to compare the efficacy and impact on QOL of two forms of pain control: continuous intravenous infusion of morphine (IV) versus patient-controlled analgesia (PCA). Pain and sedation as well as patients' psychological distress and preference for having personal control over the administration of morphine were assessed. While

the PCA group was found to have used significantly less morphine ($P<.05$) and reported significantly greater pain intensity ($P<.05$) than the IV group, there was an equivalent rating of pain relief in both arms. Furthermore, those on PCA reported significantly less psychological distress at day 5 than those in the IV arm ($P<.05$). Those on PCA reported the least sedation and had the least distress of all subgroups, controlling for other sociodemographic and pain characteristics (*9*).

CALGB-8971—A dose–response trial of megestrol acetate for the treatment of cachexia in patients with advanced lung or colorectal cancer. The objective of this study was to evaluate the effect of three different levels of megestrol acetate on weight gain and QOL of cachectic patients. The QOL component of this study was closed prior to completion as a consequence of a 75% drop-off in patient assessment by the 1st month assessment due to illness, death, and interviewer error. Although severely compromised by attrition, data analysis revealed no significant differences in QOL due to dose level of megestrol acetate, for the entire sample or by disease site.

Long-Term Psychosocial Adaptation of Cancer Survivors

Closed Protocols

CALGB-8561—Comparative assessment of psychosocial sequelae in long-term Hodgkin's disease survivors. The objective of this study was to examine the long-term psychosocial adaptation of 273 survivors of advanced Hodgkin's disease who had been treated in any of nine CALGB clinical trials. Psychological distress was found to be elevated by one standard deviation above that of healthy respondents, as assessed by the Brief Symptom Inventory, with 22% reporting distress at a level requiring further psychiatric evaluation. Furthermore, a range of psychosocial "re-entry" problems was reported as a consequence of having had Hodgkin's disease: denial of life insurance (31%) and health insurance (22%), sexual problems (37%), conditioned nausea in response to reminders of chemotherapy (39%), and a negative socioeconomic impact on their lives (36%) (*3,4*). This study established the value of the telephone interview as the method for QOL data collection in the cooperative clinical trials group and served as the foundation for the proposed telephone counseling study to improve adaptation in survivors upon completion of their oncology treatment (CALGB-9360).

CALGB-8562—Comparative assessment of psychosocial and psychosexual sequelae in three treatment regimens for advanced Hodgkin's disease (companion study to CALGB-8251): a randomized phase III trial comparing MOPP (i.e., mechlorethamine + vincristine + procarbazine + prednisone), ABVD (i.e., doxorubicin + bleomycin + vincristine + dacarbazine), and MOPP alternating with ABVD in treatment of advanced Hodgkin's disease. CALGB-8562 was a subset of CALGB-8561, involving 92 patients who had been treated for advanced Hodgkin's disease in one of the nine clinical trials, CALGB-8251. This study was undertaken to determine if there were significant differences in survivors' long-term psychosocial and psychosexual function as a consequence of differential gonadal damage from the three regimens, MOPP versus ABVD versus MOPP/ABVD. No significant long-term advantage was found for survivors of Hodgkin's disease treated by the less gonadally toxic ABVD regimen (*5*).

CALGB-8963—Psychosocial adaptation of survivors of acute leukemia. This study was developed to examine the psychosocial adjustment of survivors of acute leukemia and to identify factors predictive of current distress. Initial analyses indicate that 14% of leukemia survivors reported psychological distress that was at a level requiring further psychiatric evaluation, as measured by the Brief Symptom Inventory. Survivors most likely to have heightened distress were younger ($P<.05$) and were less educated ($P<.002$), had a history of conditioned anticipatory distress prior to their chemotherapy treatment ($P<.05$), and had a worse family environment in conjunction with more medical problems subsequent to completion of treatment ($P<.05$) (*11*).

Psychosocial and Socioeconomic Factors as Predictors of Survival

Active Protocols

CALGB-9364—Effect of bereavement on disease recurrence and death in women with stage II breast cancer (companion study to CALGB-8541). This companion study is designed to determine whether bereavement (defined as the loss of a spouse or child) in stage II breast cancer patients subsequent to adjuvant treatment on CALGB-8541 is associated with an increased risk of recurrence or death due to breast cancer. A secondary objective is to examine the relationship of stressful life events to current psychological status, as mediated by sociodemographic, medical, and social support factors. With a case–control research design, case subjects include patients who have had disease recurrence or have died subsequent to treatment completion of CALGB-8541; control subjects are women who are alive without disease recurrence. Case and control respondents will be matched for age, menopausal status, time of entry to CALGB-8541, lymph node status, and family status (living children versus no living children).

Closed Protocols

CALGB-7761—A study to determine the effectiveness of single versus multiple alkylating agents with or without doxorubicin in the primary treatment of multiple myeloma. Psychosocial status at protocol entry was examined as a predictor of survival. Multiple myeloma patients were administered two measures of psychological state at base line: the POMS and the Multiple Affective Adjective Checklist (MAACL). Neither POMS nor the MAACL was a significant predictor of survival (*8*).

CALGB-8082—Surgical adjuvant chemotherapy for breast carcinoma: two CMFVP regimens (i.e., cyclophosphamide + methotrexate + fluorouracil + vincristine + prednisone) with or without a subsequent doxorubicin combination. This study also examines psychosocial status at protocol entry as a predictor of survival. Women with stage II breast cancer were administered the SCL-90 (Symptom Checklist, a measure of psychological state) at base line. After control-

ling for known medical prognostic factors, the SCL-90 total score was not found to significantly predict survival at 7 years.

Future Plans

Interventions to Improve Patient Adaptation

CALGB-9360—Psycho-educational/interpersonal counseling intervention to improve "re-entry" adjustment of Hodgkin's disease patients upon completing active treatment: a pilot study (companion study to CALGB-8952). The major objective of this proposed study is to evaluate the feasibility of conducting a psychosocial intervention by telephone. The intervention, consisting of education, counseling, and emotional support, improves Hodgkin's disease patients' adjustment upon treatment completion. This intervention is based on Interpersonal Counseling developed by Klerman and colleagues (*14,15*), with an expanded psycho-educational component, and has been adapted for a cancer patient population. This study will be a companion to CALGB-8952, in which patients with advanced Hodgkin's disease are randomly assigned to receive either MOPP/ABV (i.e., doxorubicin + bleomycin + vinblastine) or ABVD. Patients who completed treatment on CALGB-8952 within the past 4 months will be eligible to participate. The intervention will consist of six telephone counseling sessions, conducted biweekly by an oncology nurse. Typical problematic areas identified by Hodgkin's disease patients upon completing treatment will be discussed. Relevant educational materials will also be distributed.

CALGB-9363—"ProtoCall": a randomized trial of a telephone-based supportive/educational intervention to improve QOL, satisfaction with care, and compliance. This study will test the hypothesis that cancer patients, randomly assigned to receive a supportive counseling intervention provided over the telephone by a research nurse during active treatment, experience an improvement in their psychological and social functioning and compliance to treatment, compared with a control group not receiving the intervention. Interpersonal Counseling, the therapeutic model upon which this intervention is based, was developed by Klerman and colleagues (*14,15*) and has been adapted for a cancer patient population. Patients will be assessed by use of standardized measures of QOL.

QOL of Patients During Active Treatment

CALGB-9480—A phase III study of three different doses of suramin administered with a fixed dose schedule in patients with advanced prostate cancer. A QOL component to a dose–response trial of suramin has been drafted.

CALGB-9481—A phase III study of hepatic artery floxuridine, leucovorin, and dexamethasone versus systemic fluorouracil and leucovorin as treatment for hepatic metastases from colorectal cancer. The QOL component of this phase III trial has been developed in conjunction with the newly created Clinical Economics Working Group, which will conduct a cost analysis for this study.

Study of sexual function in postmenopausal women treated with tamoxifen. A pilot study was conducted of 67 postmenopausal women with early stage breast cancer treated by tamoxifen to examine the magnitude of tamoxifen's effect on sexual functioning (*16*). Patients were assessed for drug side effects, sexual functioning, and depressed mood by use of questionnaires and vaginal and Pap smears. Data analysis began in March 1995. The next step concerning this line of research will be discussed after analysis of the pilot data.

Long-Term Adaptation of Cancer Survivors

In development—Long-term psychosocial adaptation in breast cancer survivors treated with adjuvant therapy (companion study to CALGB-7581). The long-term psychosocial adaptation of 200 breast cancer survivors treated 15-20 years ago with the adjuvant therapy on CALGB-7581 will be studied. Survivors will be interviewed concerning their current psychological, social, sexual, and vocational functioning; breast cancer detection behaviors; and problems they attributed to having been treated for cancer. An identical battery of measures that we have used in our Hodgkin's disease survivor studies (CALGB-8561/8562) and acute leukemia survivor study (CALGB-8963), supplemented by appropriate measures for this patient population, will be used. All patients will be interviewed by telephone. Any patient in significant distress will be further evaluated by a psychiatrist via telephone interview and referred for treatment in her community.

Quality of Life of Those at High Risk for Cancer

In collaboration with the Cancer Control Committee: in development—9X6Q psychological consequences of colorectal cancer screening in first-degree relatives of colorectal cancer patients (companion study to CALGB-9173). The proposed study will serve as a companion to the colorectal screening trial of first-degree relatives of colorectal cancer patients (CALGB-9173). Colon cancer patients will be randomly assigned to receive either a direct letter from the physician sent to the patients' relatives concerning screening recommendations, or a flyer concerning screening recommendations given to the patient to be sent to their relatives. High-risk relatives will be interviewed by telephone concerning psychological distress subsequent to an evaluation of compliance to screening recommendations.

References

(*1*) Ahles TA, Silberfarb PM, Rundle AC, Holland JC, Kornblith AB, Canellos GP, et al. Quality of life in patients with limited small-cell carcinoma of the lung receiving chemotherapy with or without radiation therapy, for the Cancer and Leukemia Group B. Psychother Psychosom 1994;62:193-9.

(*2*) Janov AJ, Anderson J, Cella DF, Zuckerman E, Kornblith AB, Holland JC, et al. Pregnancy outcome in survivors of advanced Hodgkin's disease. Cancer 1992;70:688-92.

(*3*) Kornblith AB, Anderson J, Cella DF, Tross S, Henderson ES, Weiss RB, et al. Quality of life assessment of Hodgkin's disease survivors: a model for cooperative clinical trials. Oncology 1990;4:93-101.

(*4*) Kornblith AB, Anderson J, Cella DF, Tross S, Zuckerman E, Cherin E, et al. Hodgkin's disease survivors at increased risk for problems in psychosocial adaptation. Cancer 1992;70:2214-24.

(*5*) Kornblith AB, Anderson J, Cella DF, Tross S, Zuckerman E, Cherin E, et al. Comparison of psychosocial adaptation and sexual function of survivors of advanced Hodgkin's disease treated by MOPP, ABVD, or MOPP alternating with ABVD. Cancer 1992;70:2508-16.

(*6*) Kornblith AB, Hollis D, Zuckerman E, Lyss AP, Canellos GP, Cooper MR, et al. Effect of megestrol acetate upon quality of life in a dose-

response trial in women with advanced breast cancer. J Clin Oncol 1993; 11:2081-9.

(*7*) Kosty MP, Fleishman SB, Herndon JE, Coughlin K, Kornblith AB, Scalzo A, et al. Cisplatin, vinblastine, and hydrazine sulfate in advanced, non-small-lung cancer: a randomized placebo-controlled, double-blind phase III study of the Cancer and Leukemia Group B. J Clin Oncol 1994; 12:1113-20.

(*8*) Silberfarb PM, Anderson KM, Rundle AC, Holland JC, Cooper MR, McIntyre OR. Mood and clinical status in patients with multiple myeloma. J Clin Oncol 1991;9:2219-24.

(*9*) Citron M, Conaway M, Zhukovsky D, Kornblith AB, Berkowitz I, Pascall V, et al. Efficacy of patient-controlled analgesia (PCA) vs. continuous intravenous morphine (CIVM) for the treatment of severe cancer pain: CALGB 8872. Proc ASCO 1993;12:abstr 1494.

(*10*) Fleishman SB, Kosty M, Herndon J, Kornblith AB, Duggan D, Morris J, et al. Quality of life (QOL) predicts survival in advanced non-small cell lung cancer. A Cancer and Leukemia Group B (CALGB) study. Proc ASCO 1994;13:abstr 1479.

(*11*) Greenberg DB, Herndon JE, Kornblith AB, Zuckerman E, Schiffer CA, Weiss RB, et al. Long-term psychosocial adaptation of survivors of acute leukemia. Proc ASCO 1995;14:abstr 1668.

(*12*) Holland JC, Kornblith AB, Zuckerman E. A centralized model for quality of life (QOL) data collection by telephone interview in multicenter clinical trials. Proc ASCO 1992;11:abstr 1421.

(*13*) Holland JC, Herndon J, Kornblith AB, Cella DF, Cooper MR, Green M, et al. A sociodemographic data collection model for cooperative clinical trials. Proc ASCO 1992;11:abstr 445.

(*14*) Klerman GL, Budman S, Berwick D, Weissman MM, Damico-White J, Demby A, et al. Efficacy of a brief psychosocial intervention for symptoms of stress and distress among patients in primary care. Med Care 1987;25:1078-88.

(*15*) Klerman GL, Weissman MM. Interpersonal psychotherapy. In: Paykel ES, editor. Handbook of affective disorders, 2nd ed. London: Churchill Livingston, 1992:501-10.

(*16*) Mortimer JE, Knapp D, Fracasso PM, Rowland JH, Kornblith AB. Assessment of sexual function in women on tamoxifen. Proc ASCO 1994; 13:abstr 1554.

Eastern Cooperative Oncology Group (ECOG)

Diane L. Fairclough, David F. Cella

History

The Outcomes Subcommittee (formerly Quality of Life Subcommittee) is one of five subcommittees of the Health Practices Committee of the ECOG. It was established in 1990 to oversee the scientific integrity of quality-of-life (QOL) research activity within the group. Its core membership is comprised of social scientists, physicians, nurses, and statisticians. Its initial role was to stimulate and promote high-quality QOL investigations in selected clinical trials. That role quickly shifted to include quality assurance of QOL data collection efforts and has more recently emphasized scientific prioritization of study proposals, because demand for QOL research within the group has outstripped the resource availability. The future of the expanded Outcomes Subcommittee promises to be very exciting as the outcomes field matures scientifically. The following is a brief description of events and activities that explain the evolution of QOL activity and priorities within the ECOG.

The first ECOG QOL study predates the formation of the QOL subcommittee. It was initiated as a pilot feasibility study for patients with metastatic non–small-cell lung cancer in 1983 (*1*). A separate QOL pilot protocol (E4983-Assessment of Quality of Life in ECOG Patients) was written to accompany the primary therapeutic study (E1583-phase II-III Chemotherapy of Metastatic Non–Small-Cell Bronchogenic Carcinoma) using the FLIC (Functional Living Index-Cancer). Compliance to the QOL assessments dropped rapidly to 33% of survivors by 6 months. Anecdotal reports suggested that medical staff were reluctant to administer the questionnaire to seriously ill patients and that in future studies efforts to address compliance should include both patients and staff.

In 1991, as a result of the early experience in the first pilot, the QOL subcommittee began actively addressing the compliance issues by sponsoring QOL data management training at each semiannual ECOG meeting, by producing a training video addressing collection, and by initiating a centralized quality assurance program for QOL assessments within ECOG. As a result of these activities, the overall compliance in all studies activated since 1991 is approximately 85%.

Building on the previous experience, a second QOL study (C0190-Quality of Life on Breast Cancer Adjuvant Trials) was developed in a group of patients with a more favorable prognosis (E3189-Phase III Comparison of Cyclophosphamide, Doxorubicin, and Fluorouracil [CAF] and a 16-week Multi-Drug Regimen as Adjuvant Therapy for Patients with Hormone Receptor Negative: Node-Positive Breast Cancer) and limited the number of assessments to three (before, during, and after therapy). In addition, reasons for missing and incomplete assessments were prospectively monitored. Compliance (defined as a completed questionnaire) was considerably improved, dropping only from 98% to 93% over the three assessments (*2*). Notably, only 1% of all assessments were missing because of patient refusal and 4% were missing for other reasons. Half of these missing assessments occurred in patients who discontinued therapy early, demonstrating the need for explicit instruction for assessment of patients who have stopped therapy early or experienced disease progression. The majority of missing items were the result of the failure to copy both sides of the form or random skipping of the back side of two-sided forms. As a result of this experience, all QOL instruments are now distributed as one-sided copies.

The conclusions of the second QOL study described above were that both the CAF and multidrug regimens have a significant impact on QOL during therapy where the magnitude of the change in Breast Cancer Questionnaire (BCQ) scores is roughly equivalent to the pretreatment difference between patients with ECOG performance status scores of 0 and 1, and by 4 months post-treatment BCQ scores on both arms recover to pretreatment levels (*3*). The impact on QOL of the shorter but more intensive 16-week multidrug regimen is greater than the 24-week CAF regimen; however, this impact seems justified by the improved disease-free (70% versus 64%) and overall (80% versus 73%) survival at 3 years for the multidrug arm (*4*). Finally, data from the BCQ complements Common Toxicity Criteria (CTC) data. The only significant treatment difference in CTC toxicity was stomatitis [20% versus 9% Grade III and IV for the 16-week multidrug versus CAF regimen (*4*)]; however, there was no difference in the related BCQ item. In contrast, the BCQ identified an additional, clinically relevant treatment difference related to fatigue.

Rapid Growth

Because of efforts of the QOL subcommittee and the nationwide increased interest in QOL research within the cancer treatment community, the number of studies with QOL components increased dramatically over time from two active protocols in 1991 to nine in 1994. There has also been a corresponding increase in the number of patients and scheduled assessments that have more than doubled every year. There are currently (July 1995) eight active and three proposed ECOG-coordinated studies with QOL as a primary or secondary end point (Table 2). ECOG also currently participates in five other intergroup clinical studies that include a QOL component.

Future Directions

New Study Development and Prioritization

With the increasing number of active and proposed QOL studies, there is a need to focus the resources of ECOG on

Table 2. ECOG coordinated studies

Year	Active studies	Scheduled assessments
1991	2	81
1992	2	328
1993	5	962
1994	9	2000

studies where the QOL component will have a substantial impact on clinical practice. With this in mind, a QOL scientific review form has been developed with specific questions about the potential impact of the QOL results on treatment in the community or on future trials. There also has been an attempt to incorporate practical considerations into study design. For example, the length of follow-up is limited to 5 years and accrual to the QOL component of the trial is limited to the first half of the enrolled patients in a large prostate cancer trial (E7892-A Phase III Randomized, Double-Blind Trial of Adjuvant Hormonal Therapy for Surgically Treated Pathologic Stage C Carcinoma of the Prostate.) Similarly, the QOL component of a large breast cancer trial (E3193-Phase III Comparison of Tamoxifen versus Tamoxifen with Ovarian Ablation in Premenopausal Women with Axillary Node-Negative-Receptor Positive Breast Cancer) is limited to the first 367 of a total of 1684 patients. Certain diagnoses have been targeted for QOL evaluations, such as lung cancer, breast cancer, and Kaposi's sarcoma within the AIDS-related malignancies. These three disease priorities were chosen in 1992 because, at that time, they represented cancer diagnoses in which QOL was recognized as a significant issue to balance with treatment response and toxicity and because of the high degree of interest and support within those disease-oriented committees for QOL research. Since 1992, there has been considerable interest in QOL research from many other disease-oriented committees. Most notable is the Genitourinary Committee that currently leads or substantially participates in three QOL protocols.

Compliance

A target of 90% compliance for QOL assessments and 100% documentation of the reasons for mistimed or missing assessments has been set. The current rate of compliance is estimated at 85%, up from a base-line rate of approximately 70%. The improvement is the result of the accumulation of experience by local data managers and an extensive QOL training and data-monitoring initiative. Prospective documentation of reasons for mistimed or missing QOL assessments are now included in all studies. In addition to past efforts, monitoring and provision of feedback on compliance to individual institutions and affiliates have begun. One component of this monitoring is an annual award to the institutional data manager with the best record of compliance. The award includes funding to attend an ECOG meeting where the data manager will make a short presentation to the QOL data management training session.

Areas of Investigation

Completed Studies

QOL assessments are complete for the adjuvant breast study (C0190), and the results were presented at the 1995 American Society of Clinical Oncology meetings (*3*). E5592-Phase III Trial Comparing Etoposide/Cisplatin versus Taxol/Cisplatin/G-CSF versus Taxol/Cisplatin in Advanced Non-Small Cell Lung Cancer has recently completed accrual, and all QOL assessments are scheduled to be completed in the summer of 1995; the final analysis of the primary outcome data is scheduled for 1996, at which time analysis of the QOL component will be completed.

Ancillary Investigations

In addition to the treatment comparisons within each of the clinical trials, there are numerous methodologic questions of interest to ECOG including:

1) Relationship of QOL and toxicity. Does QOL provide information that toxicity data alone cannot? What toxic effects have the greatest impact on QOL (*5*)?

2) Missing item in multi-item scales: What is the best method for handling assessments with missing items?

3) Analysis methods in the presence of missing assessments: What methods are practical for analysis of the QOL studies with missing assessments due to disease and treatment-related morbidity and mortality (*6*)?

4) Cross-cultural and multilingual validation of QOL instruments in clinical trials (*7,8*).

5) Testing the equivalence of commonly used QOL instruments to allow for the possibility of better comparison of data across trials and improved communication about QOL among health care professionals (*9*).

Economic (Cost) Outcomes

As a long-term objective, ECOG investigators are interested in the integration of QOL information into decision making at the levels of both individual clinical practice and health policy. Toward that end, the ECOG has expanded the scope of scientific inquiry to include economic outcomes that bear on the overall determination of the value of a given treatment in a given cohort of patients. Cost of treatment, patient preferences for various treatments, and patient values for health status outcomes of various treatments ("utilities") are all of interest. To reflect the expanded scope, the name of the Quality of Life Subcommittee was changed in late 1994 to the Outcomes Subcommittee.

References

(*1*) Finkelstein DM, Cassileth BR, Bonomi PD, Ruckdeschel JC, Exdinli EZ, Wolfer JM. A pilot study of the Functional Living Index—Cancer (FLIC) Scale: the assessment of quality of life for metastatic lung cancer patients. An Eastern Cooperative Oncology Group study. Am J Clin Oncol 1988; 11:630-3.

(*2*) Fetting J, Fairclough D, Gonin R, et al. Compliance with a quality of life evaluation in a cooperative group trial. Proc ASCO 1994;113:abstr 1572.

(*3*) Fairclough DL, Fetting J, Cella D, Wonson W, Gonin R, Grove-Conrad M, et al. Quality of life on a breast cancer adjuvant trial comparing CAF with a 16-week regimen. Proc ASCO 1995;114:abstr 890.

(4) Fetting J, Gray R, Abeloff M, et al. CAF versus a week multidrug regimen as adjuvant therapy for receptor-negative, node positive breast cancer: an intergroup study. Proc ASCO 1995;114:abstr 83.

(5) Cella DF. Quality of life: concepts and definitions. J Pain Symptom Management 1994;9:186-92.

(6) Fairclough DL, Gelber R. Quality of life: statistical issues and analysis. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. 2d ed. New York: Raven Press, 1996.

(7) Cella DF, Lloyd SR. Data collection strategies for patient-reported information. Quality Management in Health Care 1994;2:28-35.

(8) Cella DF, Wiklund I, Shumaker A, Aaronson N. Integrating health-related quality of life into cross-national trials. Quality of Life Res 1994;2:433-40.

(9) Gonin R, Lloyd S, Cella D. Establishing equivalence between scaled measures of quality of life. Quality of Life Research. In press.

Gynecologic Oncology Group (GOG)

Donald G. Gallup, David F. Cella

History

The GOG is the only national multicenter clinical trials group devoted specifically to the treatment of cancer in women. Its primary cancer treatment committees include those for cancer of the ovary, uterine corpus, and cervix and vulva. Quality-of-life (QOL) considerations are important to the treatment of gynecologic malignancies for a number of reasons. First, in early stage disease, choices often must be made between very different treatment modalities (e.g., surgery versus radiation therapy), where traditional clinical outcomes may be equivalent or close to equivalent yet there may be a dramatic difference in acute and long-term effects. Second, in advanced disease, a treatment may have limited or no benefit to survival time and yet may improve the quality of that time by virtue of tumor-burden relief. The GOG currently has studies looking at QOL in both early stage disease and advanced disease phase III protocols. The Quality of Life Committee within the GOG is, however, young and funded only through the National Cancer Institute (NCI) Cancer Therapy Evaluation Program. Therefore, the number of active protocols is kept low to reserve resource use for only the highest priority studies within the group. The GOG has a governing structure through which scientific prioritization of protocols is accomplished by the Protocol Committee, after it receives recommendations from multidisciplinary committees, such as the Quality of Life Committee.

Committee Membership

The Quality of Life Committee is a multidisciplinary committee comprising gynecologic oncologists, nurses, statisticians, psychologists, radiation oncologists, and medical oncologists. There are currently 18 voting members on the committee.

Scientific Priorities

The Quality of Life Committee of the GOG has selected two general areas for scientific priority. The first is the area of phase III clinical trials; the second is the area of delayed effects of the treatment of curable cancers. The committee has excluded phase I and phase II trials from consideration of QOL evaluation. Until now, the committee has placed less emphasis on symptom control studies. However, now that the GOG has been funded as a Cancer Control Research Base by the Division of Cancer Prevention and Control, symptom control is likely to take on more importance. Symptom control studies will be handled by the newly formed Cancer Prevention and Control Committee.

QOL Protocols

Active

There are currently four active QOL studies in the GOG. Since the committee has been in existence for only 2 years, there are no closed or completed protocols. The four active protocols are:

1) Protocol 147: Whole abdominal radiation therapy versus combination doxorubicin-cisplatin chemotherapy in advanced endometrial carcinoma (Treatment Study 122).

This first QOL protocol in the GOG was activated as a companion protocol; however, it was changed to an integrated protocol to enhance accrual to the QOL component, which has been lagging behind accrual to the parent treatment study.

2) Protocol 152: Phase III randomized study of cisplatin and Taxol (paclitaxel) with interval secondary cytoreduction versus cisplatin and paclitaxel in patients with suboptimal stage III and stage IV epithelial ovarian carcinoma.

The purpose of this study is to evaluate the value of secondary debulking surgery in patients with suboptimal ovarian cancer. It is unclear whether this surgery improves survival time, but it remains possible that it improves the quality of survival by decreasing tumor burden and associated symptoms. The purpose of the QOL study is to contrast the relief of symptoms and improvement of QOL associated with tumor debulking with the short-term disability caused by the surgery itself.

3) Protocol 9102: Effect of alopecia on cancer patient body image and the role of audiovisual information on body image.

This study is open to only a limited number of institutions.

4) Protocol 145: Randomized study of surgery versus surgery plus vulvar radiation in the management of poor-prognosis primary vulvar cancer and of radiation versus radiation and chemotherapy for positive inguinal nodes.

Proposed

1) Protocol 137: Randomized trial of estrogen replacement therapy versus no estrogen replacement in women with stage I or stage II endometrial adenocarcinoma.

This study is somewhat controversial because of the concern about the potential carcinogenicity of hormone-replacement therapy in women with endometrial cancer. Discussion regarding sample size and appropriate end points is ongoing among the GOG, the NCI, and the U.S. Food and Drug Administration.

2) Late effects of therapy for germ cell tumor survivors.

This protocol represents the GOG's first initiative into the area of studying late medical and psychological effects of curative cancer therapies. Because ovarian germ cell tumors are rather rare, a multicenter group such as the GOG is probably the only forum in which questions of late effects for this disease can

be addressed with sufficient sample size. Although there have been a considerable number of post-treatment cancer survivor studies in diseases such as leukemia, Hodgkin's disease, and testicular cancer, there exist no comparable data for women who have been previously treated for germ cell tumors. The research protocol and questionnaire packet have been approved, and study activation is pending due to the need for external funding.

3) Protocol 99RB: Phase III randomized clinical trials with laparoscopy pelvic and periaortic node sampling, vaginal hysterectomy, and bilateral salpingo-oophorectomy (BSO) versus open laparotomy with pelvic and periaortic node sampling and abdominal hysterectomy and BSO in endometrial carcinoma clinical stage I, IA, grades I, II, and III.

After surgeons become proficient in laparoscopy staging, the purpose of this phase III trial will be to demonstrate the clinical equivalence of laparoscopy compared with open laparotomy. The QOL study is then pivotal in demonstrating that laparoscopy-assisted staging is superior by virtue of more rapid return to normal function and fewer problems with psychological well-being and body image during the short-term recovery period after surgery. This is an approved study that awaits completion of the surgical proficiency stage of the project.

Future Plans

The Quality of Life Committee will continue to place emphasis on phase III trials and late treatment effects. Two priority areas for further investigation include cervical cancer (early stage disease) and bone marrow transplantation in ovarian cancer. Because this is a relatively new committee in the GOG, there are no mature data from which to generate publications.

North Central Cancer Treatment Group (NCCTG)

Charles L. Loprinzi

History

While it can be argued that many of the cancer treatment trials in the adjuvant setting and in the advanced disease setting are indirectly related to improving the quality of life (QOL) of our patients (increasing disease-free survival time without recurrent cancer and shrinkage of metastatic cancer may improve the QOL of patients), it is generally agreed that these studies are not QOL studies per se. Nonetheless, the NCCTG does have a large program dealing with symptom control trials, which we feel are more directly related to the QOL of our patients. These trials are aimed at controlling symptoms that come about from cancer and/or from cancer therapy. They are not designed to look at the quantity of life or shrinkage of cancers but, rather, are designed to look at means to decrease bothersome symptoms, and, through this mechanism, improve the QOL of the patients being studied. Past, present, and future NCCTG research related to symptom control trials include studies aimed at 1) preventing or alleviating mucositis, 2) treatment of cancer anorexia and cachexia, 3) therapy of menopausal symptoms in patients where estrogen treatment is contraindicated, and 4) improving our ability to care for patients suffering from pain.

In 1986, a protocol was developed to study whether an allopurinol mouthwash could prevent fluorouracil (5-FU)-induced stomatitis (NCCTG-86-46-51), based on promising pilot information obtained elsewhere. This study clearly demonstrated that the allopurinol mouthwash was not useful in this situation (*1*). Subsequently, another trial (NCCTG-88-92-53) was able to clearly demonstrate that oral cryotherapy could markedly reduce 5-FU-induced mucositis (*2*). A follow-up trial (NCCTG-89-92-58) was developed to evaluate different durations of oral cryotherapy in patients receiving bolus 5-FU-based chemotherapy. The results from this protocol did not suggest any advantage for continuing oral cryotherapy longer than 30 minutes (*3*). Another protocol was developed to determine whether a chamomile preparation will be able to further ameliorate 5-FU-induced stomatitis (NCCTG-90-92-56). The data from this study did not suggest any benefit from chamomile (*4*). An additional protocol (NCCTG-90-92-53) was developed to evaluate chlorhexidine and an oral nonabsorbable antibiotic lozenge to determine whether either will be helpful in alleviating stomatitis resulting from irradiation of the oral mucosa (*5*). Based on an interim analysis, the chlorhexidine arm was closed (due to lack of benefit) while the antibiotic lozenge arm versus a placebo arm is being analyzed. Two protocols were developed to evaluate whether sucralfate can 1) inhibit 5-FU-induced mucositis (NCCTG-92-92-51), or 2) inhibit treatment-induced esophagitis (NCCTG-92-94-51). Both of these trials rapidly accrued patients and both are closed and being analyzed. Also related to treatment of therapy-related gastrointestinal mucosal injury, a protocol was opened to study whether osalazine can inhibit radiation-induced diarrhea (NCCTG-91-92-53). This trial was closed early because of excessive drug toxicity (*6*). Currently, concepts approved by the National Cancer Institute (NCI) include 1) studying antibiotic lozenges for treatment of 5-FU-induced mucositis, 2) studying sucalfate for prevention of diarrhea in patients receiving pelvic radiation therapy, 3) studying glutamine for preventing 5-FU-induced mucositis, and 4) studying glutamine for preventing radiation-induced mucosal injury.

Anorexia and Cachexia

Another active area for the NCCTG Cancer Control Program has involved studies aimed at the treatment of cancer anorexia and cachexia. After an initial trial (NCCTG-87-92-51), Kardinal et al. (*7*) suggested that cyproheptadine was not very useful in this situation. A follow-up protocol clearly demonstrated that megestrol acetate could stimulate the appetite of, and cause weight gain in, patients with severe cancer anorexia and cachexia (NCCTG-88-92-51) (*8*). The results of this trial attracted substantial national interest. Accrual was subsequently completed, with 343 eligible patients being entered in another protocol that evaluated various doses of megestrol acetate and determined that there was a positive dose–response relationship for this drug for patients with cancer anorexia and cachexia (NCCTG-89-92-55) (*9*). Another trial (NCCTG-91-92-54) determined that the drug, pentoxifylline, was not helpful to alleviating cancer anorexia and cachexia (*10*). Currently, a protocol is open to compare megestrol acetate to dexamethasone and to fluoxymesterone (NCCTG-91-92-52) in patients with cancer anorexia and cachexia. In addition, a recently closed clinical trial (NCCTG-89-20-51) was designed to determine whether megestrol acetate will improve the survival of previously untreated small-cell lung cancer patients (*11*). Two other related trials evaluated a drug that has been purported to have nutritional-enhancing properties, hydrazine sulfate, in patients with 5-FU-resistant advanced colorectal cancer (NCCTG-89-49-51) (*12*) and in patients with lung cancer receiving concomitant chemotherapy (NCCT-89-24-51) (*13*).

Menopausal Symptoms

Hot flashes can be a major problem in postmenopausal women and in male patients who have had a bilateral orchiectomy, especially since estrogen therapy is relatively contraindicated in both situations. We completed accrual on a protocol

(NCCTG-89-92-54) designed to evaluate the use of the antihypertensive medication, clonidine, in this disorder (*14,15*). Subsequently, another protocol was opened to evaluate low doses of megestrol acetate for the therapy for this problematic symptom in these patient populations (NCCTG-90-92-55) (*16*). A concept has been approved by the NCI to study vitamin E in breast cancer patients with hot flashes. Another concept has been submitted to the NCI for studying low-dose androgen therapy for symptomatic hot flashes. Another quite bothersome situation for some estrogen-deprived women is vaginal dryness and/or pruritis. Estrogen creams usually will relieve this symptom, but these are relatively contraindicated in patients with breast cancer. A study is now ongoing (NCCTG-91-39-51) to evaluate a new nonhormonal agent (Replens), which has appeared to be beneficial in some women with this problem.

Analgesic Studies

Completed and published analgesic studies include 1) a placebo-controlled trial assessing the role of the psychostimulant drug, methylphenidate, in improving pain relief and general alertness in patients requiring a strong opioid drug (NCCTG-89-92-51) (*17*), and 2) a placebo-controlled trial of a topical local anesthetic cream (EMLA cream) in the management of painful percutaneous access procedures in children (NCCTG-89-92-52) (*18*).

Other Studies

We have completed the pilot phase of a protocol designed to study the efficacy of the methods of measuring QOL in patients with advanced colorectal cancer. This project was initially a part of NCCTG-89-49-51, where we were studying hydrazine sulfate in patients with 5-FU-resistant advanced colorectal cancer. Patients entered in this trial were randomly assigned to receive their QOL measured by one of four different QOL measurement instruments (Uniscale, FLIC [Functional Living Index of Cancer], Categorical Quality of Life Index, and Investigational Pictureface scale). After approximately 130 patients were entered in this hydrazine protocol, the protocol entry was stopped because of a preliminary analysis, which demonstrated no benefit for hydrazine sulfate. To complete this QOL project, a separate protocol was developed (NCCTG-93-92-51: a randomized comparison of QOL measurement tools in patients with advanced incurable colorectal cancer). However, this study was not completed because of inadequate funding to support this work.

Thus, in summary, the NCCTG has been, is, and will continue to be actively participating in research that is specifically designed to improve the QOL of patients with cancer.

References

(*1*) Loprinzi CL, Cianflone SG, Dose AM, Etzell PS, Burnham NL, Thereau TM, et al. A controlled evaluation of an allopurinol mouthwash as prophylaxis against 5-fluorouracil-induced stomatitis. Cancer 1990; 65:1879-82.

(*2*) Mahood DJ, Dose AM, Loprinzi CL, Veeder MH, Athmann LM, Therneau TM, et al. Inhibition of fluorouracil-induced stomatitis by oral cryotherapy. J Clin Oncol 1991;9:449-52.

(*3*) Rocke LK, Loprinzi CL, Lee JK, Kunselman SJ, Iverson RK, Finck G, et al. A randomized clinical trial of two different durations of oral cryotherapy for prevention of 5-fluorouracil-related stomatitis. Cancer 1993;72:2234-8.

(*4*) Fidler P, Loprinzi CL, O'Fallon JR, Michalak J, Novotny P, Hayes D, et al. A controlled evaluation of chamomile for preventing stomatitis in patients receiving 5-fluorouracil based chemotherapy: a North Central Cancer Treatment Group trial. Proc ASCO 1995;14:534.

(*5*) Foote RL, Loprinzi CL, Frank AR, O'Fallon JR, Gulavita S, Twefik HH, et al. Randomized trial of a chlorhexidine mouthwash for alleviation of radiation-induced mucositis. J Clin Oncol 1994;12:2630-3.

(*6*) Martenson J, Hylan G, Moertel C, Mailiard J, O'Fallon J, Collins R, et al. Oslazine is contraindicated during pelvic radiation therapy: results of a randomized trial. Proc ASCO 1995;14:161.

(*7*) Kardinal CG, Loprinzi CL, Schaid DJ, Hass AC, Dose AM, Athmann LM, et al. A controlled trial of cyproheptadine in cancer patients with anorexia and/or cachexia. Cancer 1990;65:2657-62.

(*8*) Loprinzi CL, Ellison NM, Schaid DJ, Krook JE, Athmann LM, Dose AM, et al. Controlled trial of megestrol acetate for the treatment of cancer anorexia and cachexia. J Natl Cancer Inst 1990;82:1127-32.

(*9*) Loprinzi CL, Michalak JC, Schaid DJ, Mailliard JA, Athmann LM, Goldberg RM, et al. Phase III evaluation of four doses of megestrol acetate as therapy for patients with cancer anorexia and/or cachexia. J Clin Oncol 1993;11:762-7.

(*10*) Goldberg RM, Loprinzi CL, Mailliard JA, O'Fallon JR, Krook JE, Ghosh C, et al. Pentoxifylline for treatment of cancer anorexia and cachexia? Randomized, double-blinded, placebo controlled trial. J Clin Oncol 1995;13:2856-9.

(*11*) Rowland KM, Loprinzi C, Shaw EG, Maksymiuk AW, Kuross SA, Jung SH, et al. Randomized double blind placebo controlled trial of cisplatin and etoposide plus megestrol acetate/placebo in extensive stage small cell lung cancer: a North Central Cancer Treatment Group study. J Clin Oncol. In press.

(*12*) Loprinzi CL, Kuross SA, O'Fallon JR, Gesme DH Jr, Gerstner JB, Rospond RM, et al. Randomized, placebo-controlled evaluation of hydrazine sulfate in patients with advanced colorectal cancer. J Clin Oncol 1994;12:1121-5.

(*13*) Loprinzi CL, Goldberg RM, Su JQ, Mailliard JA, Kuross SA, Maksymiuk AW, et al. Placebo-controlled trial of hydrazine sulfate in patients with newly diagnosed non-small cell lung cancer. J Clin Oncol 1994;12:1126-9.

(*13*) Loprinzi CL, Cianflone SG, Dose AM, Etzell PS, Burnham NL, Thereau TM, et al. A controlled evaluation of an allopurinol mouthwash as prophylaxis against 5-fluorouracil-induced stomatitis. Cancer 1990;65: 1879-82.

(*14*) Goldberg RM, Loprinzi CL, O'Fallon JR, Veeder MH, Miser AW, et al. Transdermal clonidine for ameliorating tamoxifen-induced hot flashes. J Clin Oncol 1994;12:155-8.

(*15*) Loprinzi CL, Goldberg RM, O'Fallon JR, Quella SK, Miser AW, Mynderse LA, et al. Transdermal clonidine for ameliorating postorchiectomy hot flashes. J Urol 1994;151:634-6.

(*16*) Loprinzi CL, Michalak JC, Quella SK, O'Fallon JR, Hatfield AK, Nelimark RA, et al. Megestrol acetate for the prevention of hot flashes. N Engl J Med 1994;331:347-52.

(*17*) Wilweding MB, Loprinzi CL, Mailliard JA, O'Fallon JR, Miser AW, van Haelst C, et al. A randomized, crossover evaluation of methylphenidate in cancer patients receiving strong narcotics. Support Care Cancer. In press.

(*18*) Miser AW, Goh TS, Dose AM, O'Fallon JR, Niedringhaus RD, Betcher DL, et al. Trial of a topically administered local anesthetic (EMLA Cream) for pain relief during central venous port accesses in children with cancer. J Pain Symptom Manage 1994;9:259-64.

Radiation Therapy Oncology Group (RTOG)

Todd Wasserman, Deborah Bruner, Charles Scott

History

In 1991, the RTOG Quality of Life (QOL) Subcommittee was established to oversee and facilitate QOL research. It is the commitment of the RTOG to have QOL in select phase III studies in each disease site (*1,2*). Studies are chosen where the therapeutic options most warrant a QOL investigation and where there are companion issues related to health economics.

The RTOG QOL Subcommittee has a steering group that consists of the committee chairman, vice-chairman, disease-site coordinators, statistician, RTOG protocol manager, and RTOG research associate manager. The role of this group is to provide a review of all RTOG protocols, to sign off on those studies with QOL end points, and to establish policy decisions for the Quality of Life Subcommittee. It is not an objective of the group to develop new QOL instruments, but if member institutions are interested in developing new radiation-appropriate instruments, RTOG will provide them with a research arena to test these new instruments.

The RTOG QOL initiative is separate and more global than late toxicity analysis, but there is significant interaction with the Late Effects Subcommittee (*3*). The QOL Subcommittee is involved in the testing of the Late Effects Normal Tissue Scales developed by the Late Effects Subcommittee. The RTOG is working to identify late radiation therapy effects and to evaluate interventions that diminish late effects (toxicity modifications).

As part of the educational mission within the RTOG, a QOL Procedure Manual and a patient-oriented QOL video show the value of QOL research to patients and to investigators. QOL training has been incorporated into the training session for RTOG research associates. One statistician coordinates all RTOG QOL studies to ensure consistency of design and analysis across the trials. The principal QOL researchers of RTOG institutions are nurses.

In an effort to promote greater acquisition of QOL data, the RTOG has adopted the policy of putting patients in charge of their QOL data so that they are responsible for its completeness.

RTOG QOL Research Objectives

The research objectives are to: set priorities for QOL research within the group; use existing instruments for measuring QOL in a consistent manner; develop guidelines for QOL protocol development and training of RTOG investigators, interact with the statistical unit to develop realistic end points, develop procedures for data collection based on study timepoints, and initiate and develop interventional studies in response to QOL data (*4*).

Research Issues

QOL is a multidimensional construct that must incorporate the patients' perspective within its measurement (*5*). It parallels, but is distinct from, acute and late toxicity assessment. The RTOG has also established a Late Effects Subcommittee to study toxicity measurements and scales; the QOL and Late Effects Subcommittees interact. As policy, the RTOG uses existing QOL instruments in studies rather than develop new instruments and, when QOL is determined to be a study end point, all patients will be assessed with the use of a global QOL instrument. This consistent approach to instrument selection allows investigators, data managers, and statisticians to become knowledgeable about and familiar with using the instrument(s) relevant to radiation therapy questions (*6-9*). Use of existing instruments also allows comparison of RTOG results with those currently in the literature. However, disease-specific questions are developed and added to the general questionnaire, when appropriate (*10*).

All QOL research is approved by the Quality of Life Subcommittee. One committee member is assigned to act as liaison to each disease-site committee. The priorities established by the Quality of Life Subcommittee guide the use of resources. QOL studies require extensive resources in coordination, data collection, and data analysis.

The Research Associates Committee acts as a link to the QOL Subcommittee, since its members are the resource for actually conducting the QOL research. The nurse research associates have the interest, the coordination, and the patient-interview skills needed to conduct QOL research and to participate in the research in the following ways: by serving as coordinators for the conduct of QOL studies, by developing procedures for the collection of QOL data, by training investigators and research associates in interview techniques to obtain patient consent and compliance, and by developing guidelines to reduce patient attrition.

References

(*1*) Scott CB, Stetz J. Design, analysis and data management issues in quality of life trials (QOL) within the Radiation Therapy Oncology Group (RTOG). Drug Inf J 1993;27:854-5.

(*2*) Wasserman T, McDonald A. Quality of life: the patient's endpoint. Int J Radiat Oncol Biol Phys 1995;33:965-6.

(*3*) Bruner DW, Wasserman T. The impact on quality of life by radiation late effects [editorial]. Int J Radiat Oncol Biol Phys 1995;31:1353-5.

(*4*) Scott CB, Stetz J, Bruner DW, Wasserman TH. Radiation Therapy Oncology Group quality of life assessment: design, analysis, and data management issues. Qual Life Res 1994;3:199-206.

(*5*) Bruner DW. In search of the quality in quality-of-life research [editorial]. Int J Radiat Oncol Biol Phys 1995;31:191-2.

(*6*) Choucair A, Scott C, Urtasun R, Nelson D, Coia L, Curran W. Quality of life (QOL) and neuropsychological evaluation (NSE) for patients with

malignant astrocytomas (MA). RTOG 91-14. Int J Radiat Oncol Biol Phys 1995;32:178.

(*7*) Murray KJ, Nelson DF, Isaacson S, Scott C, Fischbach AJ, Porter A, et al. Quality-adjusted survival analysis of malignant glioma. Patients treated with twice-daily radiation (RT) and carmustine: a report of Radiation Therapy Oncology Group (RTOG) 83-02. Int J Radiat Oncol Biol Phys 1993;27:207.

(*8*) Murray KJ, Nelson DF, Scott C, Fischbach AJ, Porter A, Farnan N, et al. Quality-adjusted survival analysis of malignant glioma. Patients treated with twice-daily radiation (RT) and carmustine: a report of Radiation Therapy Oncology Group (RTOG) 83-02. Int J Radiat Oncol Biol Phys 1995;31:453-9.

(*9*) Scott C, Choucair A, Urtasun R, Nelson D, Coia L, Curran W. Mini-mental status exam versus the Radiation Therapy Oncology Group's (RTOG) Neurologic Function Status Scale. Cross-validation using patients from 91-14. Int Soc Qual of Life Research, Accepted 1995.

(*10*) Watkins-Bruner D, Scott C, Lawton C, DelRowe J, Rotman M, Buswell L, et al. RTOG 90-20: a phase II trial of external beam radiation with etanidazole for locally advanced prostate cancer. Int J Radiat Oncol Biol Phys. In press.

Southwest Oncology Group (SWOG)

Laura C. Loll, Carol M. Moinpour, Polly Feigl

History

Increasing interest in cancer control research in general and effects of cancer treatment on patient quality of life (QOL) in particular led to the SWOG's first attempt at QOL assessment in SWOG-8313, an intergroup adjuvant breast cancer clinical trial. In this trial, a standard 1-year chemotherapy regimen was compared with a shorter, more intensive regimen. In 1984, the QOL study was added to the ongoing therapeutic trial but was terminated in January 1989 because of inadequate questionnaire submission rates. As compliance problems became evident, there was concern about whether QOL research could be conducted in cooperative group trials. To evaluate these concerns, the SWOG initiated a review of QOL assessment issues and methods in November 1987. A draft position paper was circulated within the SWOG outlining how, and to what extent, QOL end points should be included in SWOG clinical trials given the special needs and constraints of cooperative group research. Input from reviewers outside of the SWOG was also incorporated. In 1988, the members of the Quality of Life Subcommittee and its parent committee, the Cancer Control Research Committee, approved the QOL assessment recommendations. In April 1989, the QOL policy guidelines were approved by the SWOG's Board of Governors; the results of this review were published (*2*). The approval of the relevant committees in the SWOG and its Board of Governors was important in recognizing the legitimacy of this research in the cooperative group mechanism.

The original QOL assessment guidelines addressed a number of areas: 1) QOL assessment should occur primarily in phase III trials, although including QOL assessment in phase II trials can inform the design of future phase III trials. It is not feasible to do QOL assessment in all phase III trials, so certain types of trials have been emphasized (e.g., protocols in which the disease site is associated with poor prognosis and palliative care objectives are paramount). 2) Comprehensive assessment of QOL requires measurement of physical, emotional, and social functioning; symptom status (both disease- and treatment-related); and global perception of QOL. Symptoms associated with comorbidity should also be assessed. 3) QOL assessments should emphasize a patient report as a supplement to physician-rated toxic effects. 4) The QOL assessments should be brief questionnaires, not interviews, to reduce patient and staff burden. Example questionnaires were suggested. 5) Patient-completed QOL questionnaires should have adequate psychometric properties. 6) Categorical versus visual analogue scales are more practical for multicenter clinical trial research. 7) QOL should be assessed at least three times: before, during, and after treatment. 8) Special quality control procedures are required to monitor questionnaire submission and to enhance data quality. 9) QOL studies are conducted as companion trials to therapeutic trials, and all proposals are reviewed by the Quality of Life Subcommittee.

In 1994, the Quality of Life Subcommittee and Behavioral Sciences Subcommittee were combined in a single committee with a broad health outcomes focus. The new Behavioral and Health Outcomes Subcommittee will emphasize QOL, recruitment and adherence interventions, supportive care, and health economics; the subcommittee sees itself as a resource to the Cancer Control Research Committee and the disease committees in the SWOG.

In 1995, the QOL policy guidelines were updated to reflect the incorporation of QOL studies in therapeutic protocols versus separate companion protocols; a renewed emphasis on assessment in phase III trials; the elimination of the list of appropriate questionnaires, because the questionnaire pool is evolving; additional quality control procedures; and the structural change in the subcommittee.

Over the years, quality-control procedures evolved from responsibility at the study coordinator level to an increasing Statistical Center role. Incorporation of the QOL questionnaires in SWOG's Expectation Report, a monthly listing of overdue data by institution, and increased monitoring by the Statistical Center Data Coordinators have improved both submission rates and data quality.

Group Protocol Development Plan

Concepts can be drafted by Behavioral and Health Outcome Subcommittee members or by members of other SWOG committees. These concepts are reviewed by subcommittee members, members of the Cancer Control Research Committee, and relevant staff at the Statistical Center. If deemed to be of scientific value, which is feasible given the SWOG's structural and resource constraints and consistent with the guidelines for cancer control research, the concept is developed into a protocol by the investigator with assistance from the SWOG Statistical Center and Operations Office staff. Many levels of review occur, and the time frame from concept to protocol activation is typically more than 1 year.

QOL Protocols

Active Studies

SWOG-8994—Evaluation of QOL in patients with stage C adenocarcinoma of the prostate enrolled in SWOG-8794 (INT-0086). All patients registered in SWOG-8794 are registered in SWOG-8994 until a total of 400 patients (200 per arm) are registered to SWOG-8994. The objectives of the study are as

follows: 1) to compare three primary aspects of QOL (treatment-specific symptoms and physical and emotional functioning) according to treatment assignment in SWOG-8794, and 2) to compare three secondary aspects of QOL (general symptoms, global perception of QOL, and social functioning) according to treatment assignment in SWOG-8794. The SWOG Quality of Life Questionnaire is a battery of scales including SF-20 and SF-36 scales, the Symptom Distress Scale, and disease- and treatment-specific items. As of January 1, 1995, compliance to the submission of questionnaires has been good. The base-line QOL assessment has been submitted for 95% of the patients. The current submission rates for the 6-week, 6-month, 1-year, 2-year, and 3-year follow-up questionnaires are 89%, 91%, 88%, 75%, and 71%, respectively, for those patients alive and in the study long enough for these assessments to be made.

SWOG-9208—Health status and QOL in patients with early stage Hodgkin's disease: a companion study to SWOG-9133. It is anticipated that 288 patients will be accrued to this study before the treatment study, SWOG-9133, meets its accrual goal. The objectives of the study with respect to QOL are as follows: 1) to evaluate prospectively the health status and QOL in patients with early stage Hodgkin's disease receiving either subtotal nodal irradiation or short-course chemotherapy followed by subtotal nodal irradiation; 2) to describe the short-term, acute effects of two treatments for patients with early stage Hodgkin's disease with the use of a patient report of symptoms and QOL; and 3) to evaluate the intermediate and long-term effects of these treatments with the use of patient QOL reports over 7 years. The Symptom and Personal Information Questionnaire, the Cancer Rehabilitation Evaluation System Short Form, and the cover sheet are completed prior to registration to SWOG-9133.

SWOG-9346—A phase III trial of intermittent androgen deprivation in patients with stage D2 prostate cancer. A primary objective of this trial is to compare three treatment-specific symptoms and physical and emotional functioning by treatment arm. A secondary objective is to compare general symptoms, role functioning, global perception of QOL, and social functioning between treatment arms. This will be the first SWOG protocol where QOL is integrated into a phase III therapeutic trial. QOL is a primary objective of this trial.

Closed Studies

SWOG-9045—Evaluation of QOL in patients with advanced colorectal cancer enrolled in SWOG-8905. A total of 287 patients were registered to this QOL companion study when the parent study, SWOG-8905, closed. The objectives of this study were as follows: 1) to compare three primary aspects of QOL (treatment-specific symptoms and physical and emotional functioning) according to treatment assignment on SWOG-8905; and 2) to compare four secondary aspects of QOL (general symptoms, role functioning, global perception of QOL, and social functioning) according to treatment assignment in SWOG-8905. The SWOG Quality of Life Questionnaire was used. At trial closure, the base-line questionnaire had been submitted for 98% of the patients. The submission rates for the 6-, 11-, and 21-week follow-up questionnaires were 85%, 79%, and 79%, respectively, for those patients who were alive and in the study long enough for these assessments to have been made. The results of this study have been presented at the 1995 SWOG Fall Meeting Plenary Session.

SWOG-9039—Evaluation of QOL in patients with clinical stage D2 cancer of the prostate enrolled in SWOG-8894. A total of 739 patients were registered to SWOG-9039 when the parent study, SWOG-8894, closed. The objectives of the study were as follows: 1) to compare three primary aspects of QOL (treatment-specific symptoms and physical and emotional functioning) according to treatment assignment in SWOG-8894; and 2) to compare four secondary aspects of QOL (general symptoms, role functioning, global perception of QOL, and social functioning) according to treatment assignment in SWOG-8894. The SWOG Quality of Life Questionnaire was used. As of June 1995, 97% of the base-line QOL questionnaires had been submitted. The submission rates for the 1-, 3-, and 6-month QOL assessments were 87%, 86%, and 79%, respectively, for patients alive and in the study long enough for these assessments to have been made. Analyses are currently under way.

SWOG-9248—Phase II trial of paclitaxel (Taxol) in patients with metastatic refractory carcinoma of the breast. At study closure, 135 patients had been registered to the therapeutic protocol; of these, 18 were ineligible. One hundred twenty-five patients had completed base-line QOL questionnaires. Because of the phase II status of this trial, the QOL objective was restricted to monitoring patient reports of symptoms during treatment with paclitaxel. The Patient Symptom Monitoring Questionnaire (Symptom Distress Scale and treatment-specific items) was collected at base line and prior to each course of therapy as long as the patient remained in the protocol treatment. Analyses are currently under way.

SWOG-9235—Phase II trial of Casodex in patients with advanced prostate cancer who failed conventional hormonal manipulation. Fifty-three patients were accrued in 6.5 months prior to closure. Four patients were ineligible because of insufficient information. Because of the phase II status of this trial, the QOL objective was restricted to assessing the tolerance and toxicity of Casodex through a combination of physician and patient reporting. The Patient Symptom Monitoring Questionnaire and the McGill Pain Questionnaire were collected at base line (prestudy) and every month for 6 months, then discontinued. These data have yet to be analyzed.

SWOG-9021—Phase III study of postoperative radiotherapy for single-brain metastases. This study was closed prematurely because of poor accrual to the therapeutic portion of the trial. At the time of closure, 54 patients had been registered in the trial, 16 of whom were ineligible. The QOL objectives were to compare the two arms with respect to QOL and to evaluate the use of a QOL questionnaire specific for central nervous system malignancies. The Spitzer Quality of Life Index was filled out by the patient and a family member at each assessment. Portions of the SWOG QOL and symptom questionnaire were completed by the patient at each assessment. Concordance of patient and proxy QOL report will be examined.

SWOG-8861—Evaluation of QOL in patients with clinical stage A2 or B adenocarcinoma of the prostate enrolled in SWOG-8890. This study was closed because of poor accrual to

the therapeutic trial. The objectives of the study were as follows: 1) to compare three primary aspects of QOL (treatment-specific symptoms and physical and emotional functioning) according to treatment assignment; 2) to compare four secondary aspects of QOL (general symptoms, role and social functioning, and global perception of QOL) according to treatment assignment; and 3) to assess the feasibility of collecting QOL data from patient report via self-administered questionnaires over a 5-year period in a cooperative group setting.

Studies in Development

SWOG-9327—Randomized phase II pilot study of pentoxifylline (Trental) and placebo in patients with metastatic malignancy and the anorexia/cachexia syndrome. The objectives of this study with respect to QOL are as follows: 1) to evaluate the effect of pentoxifylline on the QOL of patients with the anorexia/cachexia syndrome related to malignancy; and 2) to evaluate the effect of pentoxifylline on the nutritional status of patients with cancer cachexia and on various laboratory measurements of nutritional status. The primary end points in this double-blinded, placebo-controlled trial are appetite and fatigue. The SWOG Quality of Life Questionnaire was modified to include a physical functioning scale more sensitive to dysfunction of end-stage cancer patients (Self-Report Barthel Index) and the Energy/Fatigue scale from the SF-36 Health Survey. The Quality of Life Questionnaire, a nutritional status form, and a pill count form completed by the SWOG institution staff will be collected. This study will be activated in early 1996.

Phase III trial of placebo versus megestrol acetate at a dose of 20 mg per day versus megestrol acetate at a dose of 40 mg per day as treatment for symptoms of ovarian failure in women treated for breast cancer (no SWOG No.). This study does not contain a comprehensive assessment of QOL but emphasizes menopausal symptoms. Patients experiencing hot flashes will be followed for 9 months. Assessment schedules and forms are under development.

SWOG-9324—Phase II trial of vinorelbine tartrate for patients with relapsed ovarian cancer. The SF-36 questionnaire and the Symptom Distress Scale will be used to describe the change in patient report of QOL (primarily symptom status) associated with salvage therapy.

Abstracts and Publications

Hayden KA, Moinpour CM, Metch B, Feigl P, O'Bryan RM, Green S, et al. Pitfalls in quality-of-life assessment: lessons from a Southwest Oncology Group breast cancer clinical trial. Oncol Nurs Forum 1993;20:1415-9.

Moinpour CM, Feigl P, Metch B, Hayden KA, Meyskens FL Jr, Crowley J. Response. J Natl Cancer Inst 1989;81:1106-7.

Moinpour CM, Hayden K, Thompson I, Feigl P, Metch B. Quality of life measurement in Southwest Oncology Group trials: policies and implementation. In: Tchekmedyian NS, Cella DF, editors. Quality of life in oncology practice and research. Williston Park (NY): Dominus Publishing Co., 1991:43-9.

Moinpour CM. Quality of life assessment in Southwest Oncology Group clinical trials: translating and validating a Spanish questionnaire. In: Quality of life assessment in health care settings. WHO/IPSEN Foundation Series. Berlin: Springer-Verlag, 1994:83-97.

Moinpour CM, Savage M, Hayden KA, Sawyers J, Upchurch C. Quality of life issues in cancer. In: Dimsdale JE, Baum A, editors. Perspectives on behavioral medicine. Hillsdale (NJ): Lawrence Erlbaum Assoc, 1995:79-95.

Thompson I, Crawford ED, Miller G, Paradelo J, Blumenstein B, Wolf M, et al. Adjuvant radiotherapy following radical prostatectomy for pathologic stage C adenocarcinoma of the prostate: initial evaluation of toxicity. Proc ASCO 1992;11:212.

Reference

(*1*) Moinpour CM, Feigl P, Metch B, Hayden KA, Meyskens FL Jr, Crowley J. Quality of life end points in cancer clinical trials. J Natl Cancer Inst 1989;81:485-95.

Childrens Cancer Group (CCG)

William E. MacLean, Jr.

History

During the past 30 years, significant advances have been made in pediatric cancer treatment as indexed by traditional study end points, i.e., disease-free survival, tumor response, and overall survival. The CCG, through its multicenter clinical trials, has been a major contributor to this success. Concurrently, the CCG has focused attention on the effects of various cancers and their treatments on children's physical health and psychosocial well-being in phase II and III therapeutic trials as well as retrospective studies of long-term survivors. These studies have included measures of physical growth, gonadal function, and cardiac and lung functions, as well as measures of neuropsychologic and behavioral functioning, employability, insurability, and educational attainment. This research has had a "toxicity" orientation for the purpose of establishing the "costs" of various treatments. These results are then used as a guide to prepare subsequent frontline protocols and to inform patients and parents of potential late effects.

These protocols include studies of acute lymphoblastic leukemia (ALL) in infants where high-dose systemic chemotherapy and intensive intrathecal therapy are used instead of cranial radiotherapy (CRT) to prevent relapse (CCG-107—intensive chemotherapy for infants with ALL; CCG-1883—treatment of newly diagnosed infants with ALL under 12 months of age); a study of children with intermediate-risk ALL who received variations of the BFM (Berlin-Frankfurt-Muenster) regimen and either CRT + intrathecal methotrexate (ITMTX) or ITMTX alone as central nervous system prophylaxis (CCG-105—studies of modifications in BFM therapy for intermediate-risk ALL; successor to CCG-162A); studies of childhood brain tumors that examine the effects of reduced radiotherapy (CCG-923—low stage medulloblastoma: a study of reduced neuraxis irradiation in newly diagnosed children; CCG-9891—low-grade astrocytoma and CCG-9892—treatment of medulloblastoma and primitive neuroectodermal tumor in children older than 36 months to 10 years of age with reduced neuraxis radiotherapy and adjuvant chemotherapy); a study of brain tumors in infants that compares two chemotherapeutic regimens in conjunction with granulocyte colony-stimulating factor (CCG-9921—multiagent chemotherapy and deferred radiotherapy in infants with malignant brain tumors); a study of bone marrow transplant (BMT) in first remission of ALL (CCG-1921—allogeneic BMT in first remission for children with high-risk features of ALL); and a retrospective study of fertility and psychosocial status in long-term survivors of childhood ALL (L-891).

Although much of the research contained in the therapeutic protocols is ongoing, several preliminary reports have been published (*1-6*). The long-term survivor study (L-891—retrospective cohort study of late effects in long-term survivors of childhood ALL) has yielded several interesting findings. For example, survivors (ages 18-33 years) scored significantly higher (more anxiety and more depression) on the Profile of Mood States (POMS) than sibling controls (*6*). Female survivors had higher scores on the POMS than did male survivors or female and male siblings. Survivors who reported unemployment because of the effects of their disease scored significantly higher on the POMS than did survivors who reported no disease-related employment problems and were fully employed. Similar effects were evident in relation to schooling. Interference with education was associated with higher POMS scores. In relation to both employment and education, the difference in scores was significantly greater for those survivors who were older at diagnosis compared with those who were younger.

These survivors were also questioned about their scholastic performance (*1*). After diagnosis, survivors were more likely than their sibling control subjects to enter a special education or learning disabilities program but just as likely to enter a program for gifted and talented children. The risk associated with special education and learning disabilities placement increased with increasing dose of cranial radiotherapy. Despite these problems, survivors generally had the same probability as their siblings of finishing high school, entering college, and earning a bachelor's degree. There was some indication that survivors treated with 24 Gy and those diagnosed before 6 years of age were less likely to enter college.

QOL Protocols

CCG currently has three protocols in varying stages of development that will include QOL end points.

CCG-1941—BMT versus prolonged intensive chemotherapy for children with ALL after an initial bone marrow relapse. This phase III trial for children with ALL and an initial bone marrow relapse within 1 year of completion of therapy will compare prolonged intensive chemotherapy, conventional bone marrow transplantation using human leukocyte antigen/mixed leukocyte culture (HLA/MLC)-compatible sibling donors, and alternative bone marrow transplant strategies employing alternative stem cell sources, e.g., matched unrelated marrow donors, haploidentical family marrow donors, or purged autologous marrow. The study plan includes health status assessments with the use of the Ontario Health Survey at several time points. Additional measures of social, emotional, and physical functioning are being considered for inclusion.

CCG-1951—Extramedullary relapse and occult marrow involvement in childhood ALL. This is a phase III group-wide study of children with ALL whose first adverse event while on or off therapy is a central nervous system (CNS) or testicular relapse. Therapy will be determined by the time and site of oc-

currence of the extramedullary relapse. Patients developing an early relapse in CNS, less than 18 months from first complete remission, who have an available HLA/MLC-compatible sibling bone marrow donor will be eligible for allogeneic BMT. For patients developing an early CNS relapse without an available HLA/MLC-compatible sibling bone marrow donor, for late CNS relapse, and for all testicular relapse patients, induction therapy will be followed by four 6-week intensification cycles of chemotherapy and by four 12-week maintenance cycles. Patients with CNS relapse will be given craniospinal irradiation during the initial month of maintenance at dosages being determined by current treatment regimen (BMT versus chemotherapy) and previous CNS radiotherapy history. The health status assessment and social, emotional, and physical functioning measures will be the same as those used in CCG-1941.

S-942—Study of minimally invasive survey of the chest in children with cancer. This is a study comparing minimally invasive surgery (MIS) with conventional open-chest surgery in the management of cancer in children. A secondary aim of the study is to evaluate the impact of MIS and open surgery on short-term QOL, at 3, 7, and 30 days after surgery. Several domains of QOL will be examined, including surgery-related pain; physical, social and emotional functioning; and global ratings of health and overall QOL.

Group Development Plan

It has been argued that QOL is not synonymous with measures of intelligence, psychopathology, academic achievement, peer social status, neuropsychologic functioning, health status, fertility, sensation, mobility, self-care, pain, or growth. Rather, QOL is defined in the literature as a multidimensional construct composed of social, emotional, and physical functioning as perceived by the patient. Unfortunately, there are few measures of QOL consistent with this definition that are appropriate for the special conditions associated with pediatric oncology. These conditions include a rapidly developing person in which functioning changes radically through the developmental age span, the need for informants or proxies for young children, the need to consider family and cultural context in assessing a particular child's QOL, measurement of generic aspects of QOL and disease- or treatment-specific effects, the need to fit with large-scale multi-institutional protocols, and so on. Simply stated, what single measure could possibly encompass all of the dimensions of QOL across a developmental age span of 18 years or more, be sensitive to changes in functioning that result from cancer and its therapy, and be suitable for use in the cooperative group research context?

Several CCG committees (e.g., Nursing, Psychology, Cancer Control, and Supportive Care) have been discussing these issues while developing a group-wide plan on QOL. The CCG Executive Committee is in the process of establishing a single strategy group that will determine research priorities for late effects, cancer control, supportive care, and QOL. This strategy group will focus its future efforts on measurement issues and several high-priority research studies.

Measurement is a primary concern for QOL research in pediatric oncology. We are using the few existing measures in current studies to gain some experience with them and to assess issues related to compliance and respondent burden. Concurrently, there is considerable interest in the development of new QOL measures appropriate for use in future protocols. Several candidate measures are being developed that warrant consideration after determining their psychometric characteristics and sensitivity to change in QOL over successive observations. In this regard, we plan a study of ALL patients that will yield important information regarding the Minneapolis–Manchester QOL measure in comparison with the currently available measures. This instrument, recently developed by M. Jenney from the U.K. and her colleagues at the University of Minnesota, is a refinement of several existing measures of health outcomes. The proposed study will provide important validity data and a demonstration of the feasibility of telephone interviewing for QOL data collection.

There are plans to conduct three retrospective studies involving three well-known patient cohorts: children with acute myelogenous leukemia who received BMT versus chemotherapy, children with brain tumors who received either standard versus reduced radiotherapy, and children with non-Hodgkin's lymphoma who received eight-drug combination chemotherapy versus those who received four-drug combination chemotherapy followed by low-dose regional radiotherapy. These studies will provide much needed data on long-term QOL in these patients.

CCG will also be examining the appropriateness of existing QOL measures for all cancers. Some have argued that the available measures are most appropriate for children with leukemia and that they are not particularly sensitive to the effects of brain tumors and their treatment. It could be that we will direct some effort to developing a brain-tumor-specific pediatric QOL measure for use in CCG studies.

References

(*1*) Haupt R, Fears TR, Robison LL, Mills JL, Nicholson HS, Zeltzer LK, et al. Educational attainment in long-term survivors of childhood acute lymphoblastic leukemia. JAMA 1994;272:1427-32.

(*2*) Kaleita TA, MacLean WE, Reaman G, Whitt JK, Hammond GD. Neurodevelopmental studies of children less than 12 months old diagnosed with acute lymphoblastic leukemia. Proc Inter Soc Pediatric Oncol 1987; 19:159.

(*3*) MacLean WE Jr, Noll RB, Stehbens JA, Kaleita TA, Schwartz E, Whitt JK, et al. Neuropsychological effects of cranial irradiation in young children with acute lymphoblastic leukemia 9 months after diagnosis. Arch Neurol 1995;52:156-60.

(*4*) Stehbens JA, MacLean WE, Kaleita TA, Noll RB, Schwartz E, Cantor N, et al. Effects of CNS prophylaxis on the neuropsychological performance of children with acute lymphoblastic leukemia: nine months post diagnosis. Child Health Care 1994;23:231-50.

(*5*) Stehbens JA, Kaleita TA, Noll RB, MacLean WE Jr, O'Brien RT, Waskerwitz MJ, et al. CNS prophylaxis of childhood leukemia: what are the long-term neurological, neuropsychological, and behavioral effects? Neuropsychol Rev 1991;2:147-77.

(*6*) Zeltzer L, Zhang F, Stuber M, Meadows A, Mills J, Byrne J, et al. Psychological sequelae in adult survivors of childhood acute lymphoblastic leukemia. Med Pediat Oncol 1994;23:169.

Pediatric Oncology Group (POG)

Andrew S. Bradlyn, Brad H. Pollock

History

The POG established a Quality of Life (QOL) Committee approximately 4 years ago, and that group currently is organized as a subcommittee of the Cancer Control Committee. To date, QOL outcomes have been included in a small number of trials. Initially, there was debate regarding a number of conceptual and methodologic issues that evolved into the development of a set of guidelines to direct research efforts. In investigating the QOL of children being treated for, or ultimately surviving, a malignant cancer, POG investigators have faced many of the typical problems that confront all investigators in this field: recruiting subjects, determining the most appropriate time of assessment for a particular protocol, and dealing with missing data. However, there has been a need to address a number of issues that are somewhat unique to children and families, such as establishing a definition of QOL that is applicable to children, adolescents, and families; dealing with the rapid and variable developmental changes that occur throughout the 0-18+ year life span of our patients; identifying instruments that reflect that QOL definition; and finally, dealing with the ever-present (and potentially paralyzing) problem of proxy respondents. Historically, we know that QOL outcomes by almost any definition have only rarely been included in phase III trials by either of the two pediatric cooperative groups (*1*); POG has made a concerted effort over the past several years to examine the potential contribution of alternate end points, such as QOL and economic factors (*2*).

Definition

The following definition of QOL was adopted by the POG on the basis of the World Health Organization's definition of health (1958):

> "Quality of life is a multidimensional construct, incorporating both objective and subjective data, including (but not limited to) the social, physical, and emotional functioning of the child and, when indicated, his/her family. QOL measurement must be sensitive to changes that occur throughout development." (Pediatric Oncology Group: unpublished definition.)

This definition provides a focus for what is meant by the term "QOL," so that the POG research efforts could be planned, coordinated, and responsive to the rigors of the scientific method. With limited resources (financial and human), the goal is to implement a standardized but flexible approach to the assessment of QOL with the use of a core set of measures along with additional QOL measures that are specific to the objectives of the protocol.

High-Priority Trials for QOL Assessment

Recognizing the limited resources that are available, certain types of protocols were identified as being of the highest priority for QOL end points, and these are consistent with those factors typically identified in the literature. For example, phase III trials were identified as being the most relevant to QOL assessment, especially trials comparing different treatment modalities or trials expected to result in therapeutic equivalence. Additionally, it is specified that trials should be expected to accrue a sufficiently large number of subjects to ensure adequate statistical power for the QOL questions.

The two areas in which the QOL Committee has focused its efforts are as follows: 1) standard measurement strategy for the measurement of QOL, including the specific instruments that are appropriate, and 2) how to deal with the issue of proxy respondents.

QOL Measurement

In terms of measurement strategy, the POG has adopted an approach that is based on the notion of a standard core group of measures that may be supplemented by other relevant modules. Group QOL Guidelines recommend the inclusion of several different types of instruments across protocols but also allows for the inclusion of protocol-specific questions. The basic strategy is to include (at a minimum) a measure of generic health status, a cancer-specific measure, and a measure of performance status. Additionally, the inclusion of several single-item global ratings of QOL and health is recommended. Unfortunately, unlike QOL investigations with adult patients where there may be multiple, standardized, psychometrically sound instruments from which to choose, QOL research in pediatrics has been severely hampered by the relative paucity of appropriate instruments. For example, at this point in time, there is only one published measure of QOL that was developed with pediatric cancer patients and their families, although there are a number of others currently being developed (*3*).

The issue of the proxy respondent is particularly problematic in investigations of pediatric populations. Because there are limitations associated with proxies, and studies have shown that patients are the best informants about their own QOL, pediatric trials present a unique challenge when devising a QOL component. It is not unusual for POG trials to identify eligible subjects as all patients under the age of 21 with a particular malignancy. Thus, we have to develop a measurement strategy for subjects who may range in age from less than 1 year to 21 years of age. This is further complicated by the fact that patients may cross previously set age ranges for particular instruments during the course of their participation.

Ongoing activities include an effort to provide information to each of the Disease Committees within the POG about how QOL questions might be identified and included in protocols under development. A list has been established of individuals at each institution who are responsible for QOL data-collection aspects of clinical trials. In addition, a manual is in preparation that addresses issues regarding the day-to-day management of the investigation, i.e., Institutional Review Board submissions and consent forms, standard administration instructions, and typical "problem situations" and solutions.

Protocols With QOL Assessment

To date, QOL measures have been included in the following POG protocols:

POG-9202—ALinC16: acute leukemia in children No. 16. This protocol, which is currently accruing subjects, includes a modified QOL component that is embedded within psychologic studies. The QOL data relate to the objective "to determine the feasibility of gathering neuropsychological data with magnetic resonance imaging and specified neuropsychological tests."

POG-9331—Intergroup low-risk medulloblastoma. The protocol also includes a modified QOL component that is embedded within psychological studies. The QOL data relate directly to the objectives. This protocol is currently accruing subjects.

POG-9485/9585—Intergroup minimal-access surgery. This protocol includes the full QOL battery as recommended in the POG Guidelines. Additionally, questions relating to respondent burden are included to further understanding of this issue. The QOL data are end points in the primary objectives of this protocol, which are "to investigate the role of minimal access surgery in terms of short-term quality of life, economic factors, and perioperative morbidity and mortality."

There are also a number of protocols in various stages of development or review that include QOL components, including studies of the effect of Enalapril in reducing cardiotoxicity from anthracycline therapy, the effects of bone marrow transplantation on QOL, and the relationship between doxorubicin infusion time and QOL. It is important to note that all of these efforts have been multidisciplinary and have originated from a variety of disease and/or discipline committees.

Group Perspective

The QOL Subcommittee has been fortunate to have the support of the leadership of the group, which has resulted in earlier identification of relevant protocols and, importantly, the administrative and statistical support that is crucial to successfully bringing research questions of this type to fruition. In fact, the mission of the POG, as described in its constitution, has recently been amended to include not only the cure of childhood cancer but also the promotion of the quality of our patient's lives. While strides have been made toward this goal during the past 4 years, QOL research is clearly in an early phase within the POG. There is a desperate need for the funding of basic research that addresses questions such as instrument development and ongoing validation. Given the relatively low-base rate of childhood cancers, many of these questions must be asked on a multi-institutional or group-wide basis, and this cannot be accomplished without financial support. POG investigators are pleased with the progress thus far and are looking forward to contributing to the growing database regarding the QOL of children and adolescents with cancer as well as their families.

References

(*1*) Bradlyn AS, Harris CV, Speith L. Quality of life assessment in pediatric cancer clinical trials: a retrospective review of phase III cooperative group investigations. Soc Sci Med 1995;41:1463-5.

(*2*) Land V, Pollock BH. Economic outcomes in clinical trials. Symposium at the Pediatric Oncology Group meeting. Chicago, IL: 1994.

(*3*) Goodwin D, Boggs S, Graham-Pole J. Development and validation of the Pediatric Oncology Quality of Life Scale. Psycholog Assess 1994;6:321-8.

Quality of Life in Clinical Cancer Trials: Experience and Perspective of the European Organization for Research and Treatment of Cancer

*Gwendoline M. Kiebert, Stein Kaasa**

Background

The European Organization for Research and Treatment of Cancer (EORTC) is an international nonprofit organization that was founded in 1962 by European cancer specialists to conduct, develop, coordinate, and stimulate research in Europe on the experimental and clinical bases of cancer treatment and related problems. The ultimate goal of the EORTC is to provide the best state-of-the-art treatment to as many cancer patients as possible in Europe. The fundamental structure of the EORTC Treatment Division is based on the input from 22 cooperative groups, who develop their clinical research through the direct input of the participating scientists. The development of this research is supervised by different committees. Research is accomplished mainly through the execution of large, prospective, randomized, multicenter cancer clinical trials. More than 2000 clinicians located in 350 medical institutions in 31 countries enter each year approximately 6000 patients in about 100 ongoing studies.

The EORTC Data Center in Brussels is the nucleus of all the clinical research. It provides an optimal and unique European infrastructure to conduct multicenter and multidisciplinary clinical trials with expertise in data management, biostatistics, medical monitoring, and quality-of-life and health economics evaluations. The EORTC Data Center is therefore concerned with all aspects of late phase II and phase III cancer clinical trials, the design and preparation of such trials, collection of data, statistical analysis, and publication of the final results, as well as quality-control procedures and legal and administrative responsibilities. Currently, more than 50 people with various scientific backgrounds are working at the Data Center. They include data managers, statisticians, medical doctors, medical fellows, nurses, economists, pharmacologists, psychologists, and computer specialists. To address these important issues, six specialty units have been created within the Data Center.

Until recently, clinicians have mainly focused their attention on the more classical aspects of evaluating cancer treatment outcomes, such as response to treatment, relapse, and (disease-free) survival. It is now increasingly recognized that quality of life is an important outcome measure in the evaluation of cost/benefit ratios of new interventions, especially when the impact of medical treatment on the length of life is expected to be small. The number of trials that include quality of life as an outcome parameter has increased rapidly during the last few years and is still increasing. Table 1 provides an overview of the EORTC trials that have included quality of life as an end point during the past 10 years.

The rapid growth of the number of studies assessing quality of life emphasized the need for a coherent policy and a standard approach to conduct this research. For this reason, a Quality of Life Unit was established at the EORTC Data Center in 1993 with financial support from the European Community. Its main objective is to stimulate, enhance, and coordinate quality of life as a treatment outcome in cancer clinical trials. In this context, the principal tasks of this unit are to establish an adequate infrastructure for the data management of quality-of-life studies; to undertake the design, collection, and analysis of quality-of-life data in EORTC clinical trials; and to generate specific quality-of-life research questions. At present, the unit consists of a psychologist (Ph.D. and head), a statistician, a part-time quality-of-life administrator, and a part-time data manager. In the near future, we hope to welcome a research fellow.

The Quality of Life Unit has a close collaboration with and builds further on the achievements of the EORTC Study Group on Quality of Life. This group was created in 1980 and from its inception has included a broad range of professionals with extensive experience in quality-of-life research. In these past years, this group has performed much research to develop sound tools to measure quality of life in cancer patients. It has developed a modular approach to the assessment of quality of life by which a core questionnaire measuring a range of physical, emotional, and social health issues is supplemented by diagnosis-specific and/or treatment-specific modules (*1*,*2*). The core questionnaire, known as the EORTC QLQ-C30, is currently available in 18 languages and is being used in more than 200 studies worldwide. It is a copyrighted instrument, and administration of the core questionnaire is handled by the Quality of Life Unit. Various modules (such as the lung, breast, head and neck, and colorectal cancer modules) have been developed or are currently being field tested [e.g., (*3*,*4*)]. For each module,

**Affiliations of authors:* G. M. Kiebert, Quality of Life Unit, European Organization for Research and Treatment of Cancer (EORTC) Data Center, Brussels, Belgium; S. Kaasa, EORTC Study Group on Quality of Life, Palliative Medicine Unit, University Hospital, Trondheim, Norway.

Correspondence to: Gwendoline M. Kiebert, Ph.D., Quality of Life Unit, EORTC Data Center, Avenue Mounier 83/11, 1200 Brussels, Belgium.

Table 1. EORTC trials (1985-1995) with quality-of-life evaluation as an end point

Year	Phase	Protocol*	No. of patients	Status
1985	III	Operable breast cancer in the elderly	413	Closed
	III	Randomized trial on dose response in radiotherapy of low-grade gliomas	379	Closed
	III	Orchidectomy versus LHRH analogue in metastatic prostate cancer	327	Closed
1986	III	Long-term QoL of adult leukemia after bone marrow transplantation versus intensive consolidation (acute myelogenous leukemia)	1057	Closed
	III	Radiotherapy versus no radiotherapy for cerebral gliomas of the adult	237	Open
	III	Estracyt versus mitomycin C in hormone escaped advanced prostate cancer	171	Closed
1987	III	Development of EORTC core QoL questionnaire for cancer patients	985	Closed
1988	III	Adjuvant trial in malignant melanoma comparing recombinant interferon alfa-2 with recombinant interferon gamma with control	755	Open
1990	III	Early versus late orchidectomy or early versus late treatment in asymptomatic nonmetastatic prostate cancer	537	Open
	III	Endocrine treatment with flutamide versus cyproterone in good-prognosis patients with prostate cancer	286	Open
	III	Orchidectomy versus orchidectomy + mitomycin C in poor-prognosis patients with metastatic prostate cancer	189	Open
1991	III	Short, intensive preoperative combination chemotherapy versus similar therapy given postoperatively in breast cancer patients	482	Open
	III	LD-ARA-C versus LD-ARA-C + GM-CSF versus LD-ARA-C + recombinant interleukin 3 for patients with myelodysplastic syndromes and high risk of developing acute leukemia	201	Open
	III	Flutamide versus prednisolone in hormone-resistant metastatic prostate cancer	114	Open
1992	II	Second-line chemotherapy with docetaxel in patients with breast cancer	83	Closed
	III	Strontium chloride versus palliative local-field radiotherapy in patients with hormone-resistant metastatic prostate cancer	70	Open
1993	III	Dose-intensive chemotherapy as primary treatment in locally advanced inflammatory breast cancer	249	Open
	III	Influence of dose intensity on survival in G-CSF-supported treatment of HIV-associated non-Hodgkin's lymphoma (high malignancy)	209	Open
	II-III	Comparison of cisplatin-based chemotherapies in NSCLC	181	Open
	II	Randomized paclitaxel versus doxorubicin as first-line chemotherapy for advanced breast cancer	230	Open
	III	Chemotherapy with or without G-CSF in operable osteosarcoma	47	Open
	III	Induction and intensive consolidation followed by bone marrow transplantation in acute myelogenous leukemia	615	Open
1994	III	Role of booster dose of postoperative radiotherapy in patients with early stage carcinomas of head and neck	12	Open
	III	Oral pamidronate versus placebo in breast cancer patients with newly diagnosed bone metastases	77	Open
	III	Prospective radiotherapy versus chemotherapy in patients with locally advanced head and neck carcinoma	53	Open
	III	Paclitaxel + platinum versus cyclophosphamide + platinum in advanced epithelial ovarian cancer	171	Open
	III	5-FU and L-leucovorin after liver or lung metastasis resection from colorectal cancer	4	Open
	III	Cisplatin + cyclophosphamide versus abdomino-pelvic irradiation in high-risk epithelial ovarian cancer	5	Open
	II-III	Cisplatin + 5-FU versus cisplatin + 5-FU with interferon alfa in metastatic pancreatic cancer	14	Open
	III	Surgery versus radiotherapy in NSCLC after response to induction chemotherapy	13	Open
1995	II	First-line iv vinorelbin and cisplatin in patients with metastatic epidermoid carcinoma of esophagus	3	Open
	III	3BEP versus 3BEP-1EP in good-prognosis germ cell cancer	13	Open
	III	Reliability and validity of QLQ-C30 (version 3.0) and head and neck cancer module	0	Open
	III	Reliability and validity of QLQ-C30 (version 3.0) and breast cancer module	0	Open

*LHRH = luteinizing hormone-releasing hormone; QoL = quality of life; LD-ARA-C = low-dose cytosine arabinoside (cytarabine); GM-CSF = granulocyte–macrophage colony-stimulating factor; G-CSF = granulocyte colony-stimulating factor; NSCLC = non-small-cell lung cancer; 5-FU = fluorouracil; iv = intravenous; HIV = human immunodeficiency virus; 3BEP = three cycles of bleomycin + etoposide + cisplatin; 3BEP-1EP = three cycles of bleomycin + etoposide + cispatin—followed by one cycle of bleomycin + etoposide.

one member of the study group is the principal investigator responsible for its development.

How Does the EORTC Integrate Quality-of-Life Questions in Clinical Trials?

There are three principal channels to integrate quality of life as an outcome measure in EORTC clinical trials.

The first channel is through training and education. Although many clinicians subscribe to the importance of quality-of-life evaluation in clinical trials, only a few have extensive knowledge and/or personal experience with quality-of-life assessments. Often there is a lack of familiarity with quality-of-life instruments as well as a lack of experience in solving practical problems in implementing quality-of-life assessments. The Quality of Life Unit tries to increase awareness and knowledge of quality-of-life issues by having sessions on quality-of-life considerations at least once during one of the cooperative groups' meetings. These meetings are held every 6 months. Many sessions have already been organized, and it is our experience that a 1-hour presentation of the basic principles about the "when," "why," "who," and "how" of quality-of-life measurements is usually sufficient to result in a lively, constructive discussion and prepares the groundwork for future collaboration with the Quality of Life Unit at the Data Center.

The second channel for integrating quality-of-life issues is through assigning liaison members from the Quality of Life Study Group to the various disease-oriented cooperative groups. For the past few years some members from the study group, with a particular interest in involvement in a specific disease site, have been appointed as liaisons to offer their expertise to the cooperative groups on how to conduct quality-of-life assess-

ments in clinical trials. In principle, the liaisons attend all meetings of their respective cooperative group. Between meetings, they can be consulted concerning quality-of-life issues. Some cooperative groups have formed a Quality of Life Subcommittee consisting of clinicians with a special interest in quality-of-life issues. During the cooperative groups' meetings, subgroup meetings are held to discuss ongoing matters. A senior member from the Study Group on Quality of Life is also a member of the EORTC Protocol Review Committee.

The third channel is the involvement of the Data Center staff at an early phase of protocol development. The existing protocol submission procedures require the early involvement of the Quality of Life Unit at two moments. The first moment is during the initial development of a new protocol. This consists of a two-page outline, which must be submitted to the Protocol Review Committee for approval of the basic idea of the study. Each two-page outline is seen and reviewed by the Quality of Life Unit before it is sent to outside referees. If this two-page outline of the protocol is approved by the Protocol Review Committee, a full protocol can be developed in which quality of life is an integral part of the study objectives. An accompanying letter to the principal investigator includes a recommendation to contact the Quality of Life Unit as soon as possible. It is explained to the investigator that this part of the full protocol must be approved by the Quality of Life Unit before the protocol can be submitted to the Protocol Review Committee for final approval. This is the second moment. The investigator is not obliged to contact the unit, but one runs the risk that a study cannot be opened for the patients' entry because this part of the protocol has not been approved by the Quality of Life Unit.

EORTC Criteria for Inclusion of Quality of Life in Clinical Trials

Phase III Studies

The general policy of the EORTC regarding the inclusion of quality-of-life issues in phase III cancer clinical trials is as follows: Theoretically, it can be a relevant end point if

- no improvement in overall, recurrence-free, or systemic disease-free survival is expected, but when significant changes or differences in (at least) one aspect of quality of life are expected;
- one treatment results in a better survival but has more toxic effects;
- the patients have an extremely poor prognosis with or without treatment;
- treatment is known to be very burdensome to patients;
- a new (invasive) treatment is to be evaluated.

If either one or a combination of these criteria applies to a proposed study, then it is up to the cooperative group in general and the principal investigator in particular to decide whether or not quality of life will be evaluated. Both the EORTC Protocol Review Committee and the Quality of Life Unit do not follow the policy of imposing quality of life as an end point. However, they can strongly advise to include it as an end point if they consider it to be a relevant issue. The EORTC adopted this policy for the following reason: Since quality of life is a relatively new field of research (and for many clinicians and institutions even an experimental field of research), it requires extra motivation on the part of the clinicians and other persons responsible for data collection before its assessment can become a fully integrated part of clinical practice. Imposing an extra workload on already overloaded personnel, who may also doubt the usefulness of evaluating quality of life, will only have a negative effect on the quality of those data. Since there are about 100 trials ongoing every year, the EORTC prefers to have a limited number of studies with good-quality data instead of a large number of studies with low-quality data. The best way to convince and motivate people with regard to the importance of quality-of-life research is by means of examples of successful and high-quality studies.

In those studies in which quality of life has been accepted as an end point, this aspect of the study is mandatory in all institutions. The only exception to this rule is for countries for which no validated translation of questionnaires in that particular language exists. In general, the ability to fill in quality-of-life assessments is one of the inclusion criteria, but refusal or missed quality-of-life evaluations are not exclusion criteria for entering a study.

If quality of life is an end point, then the protocol should provide information on the following aspects: (*a*) rationale for the inclusion of measuring quality of life as a primary or secondary outcome measure; (*b*) formulation of both the study objectives and hypotheses; (*c*) justification for the quality-of-life aspects or dimensions that will be evaluated; (*d*) description of patient eligibility criteria; (*e*) design and methods used; and (*f*) statistical considerations such as sample size calculation and methods used to analyze data.

Phase II Studies

In principle, quality of life is not considered a relevant end point in EORTC phase II trials, since the primary aim of such a study is to determine anticancer activity as well as toxicity. Previous studies (*5,6*), however, have shown that patients and their physicians can differ in their rating of toxic effects and burden of treatment. Moreover, the clinical evaluation of toxicity focuses mostly on its occurrence and severity and not on the duration of toxic symptoms or on the relative burden for patients. For these reasons, it may be important to include the patient's valuations of these factors in phase II studies. This may provide not only important information (and thus increase our understanding of the frequency, severity, and burden of the side effects), but also valuable information for deciding on the design of a subsequent phase III trial. It can provide a better indication of which aspects of quality of life should be examined, and it may be used to determine which interventions should be made early in the phase III trial in order to minimize symptoms and dysfunction. The subjective evaluation of the perceived burden of treatment-related side effects, however, is not to be confused or regarded as equivalent to the "classical" approach to the evaluation of quality of life. The latter entails more than just the assessment of treatment-related symptoms.

There is one exception to the general rule not to measure quality of life in a phase II study: randomized phase II studies

that will continue as phase III studies, in which quality of life is regarded as an important outcome measure. Since the data on patients entered in a randomized phase II study will be included in the phase III comparison, quality-of-life assessment should have started already in phase II.

The points mentioned above reflect the present policy of the EORTC, and they serve as theoretical guidelines for the integration of quality-of-life issues in its late phase II and phase III clinical trials. In practice, however, the awareness, knowledge, and previous experience with quality-of-life evaluation appear to be stronger determinants for the integration of this end point than guidelines. These factors can be active at the following three levels: personal, group, and national.

The personal level refers to the principal investigator. If this person has had positive experience with quality-of-life evaluations, then this optimizes the chances that quality of life will be evaluated in a new study if it is considered a relevant outcome.

The same principle applies to the second level, which refers to the attitude toward and experiences of the cooperative group with quality-of-life issues. It is remarkable to note that there are cooperative groups who have a long history of quality-of-life evaluations in their trials, whereas other groups appear quite reluctant and resistant to consider quality of life as an outcome measure. This discrepancy seems difficult to explain. Although there are disease sites that have a long history of extensive quality-of-life research (e.g., breast cancer and genitourinary cancers), grounds do not seem to exist for the assumption that the relevancy of quality-of-life issues is different in the various cancer sites.

The third level concerns the national policies of the various countries. EORTC studies are conducted on an international level. Cross-nationally, substantial differences do exist, not only with regard to familiarity with quality-of-life evaluations, but also with regard to the infrastructure for managing cancer clinical trials in general and for quality-of-life assessment in particular. Some countries (e.g., The Netherlands) provide data management support to their large institutions, in the form of either data managers or research nurses. The presence or absence of such an infrastructure substantially influences the motivation and capacity of clinicians and institutions to participate in high-quality and sophisticated cancer clinical trials. Lack of data management support can be a reason to limit the number of ongoing studies that include quality of life as an end point per disease site.

How Does the EORTC Build on Successive Trials?

Ideally, each new clinical trial builds on the results of the preceding study. The efficacy of a new treatment applied in an experimental group is compared with a control group who usually receives the treatment that is considered to be the current standard. The new treatment modality is tested for equivalence or for a difference. If the new treatment is proven to be better, it will become the new standard. In the next study, this treatment will then serve as the control arm. The same principle applies to quality-of-life issues. Ideally, new studies incorporate the results of the preceding ones. Obviously, this has been the case in the process of developing the EORTC core questionnaire and the disease-specific modules. The first generation of the core questionnaire, consisting of 36 questions, was developed in 1987. Detailed results on the international field testing of this instrument were published in 1991 (*7*). While the overall psychometric results were promising, they also pointed to some directions in which the questionnaire could be improved. In the next generation of the instrument, these areas were further developed.

A major advantage of the EORTC approach is that the same core instrument is used in each study. This approach allows a sufficient degree of generalizability for cross-study comparison. The Quality of Life Unit is involved in a wide range of studies across cooperative groups and throughout all its phases from the design to the analysis and publication of the results. This unique characteristic allows us to get a good overview of the actual field of research; it enables the coordination of research projects at a European level, and it is extremely helpful in generating new research questions. This way, we hope to contribute substantially by building on the results of successive trials within the EORTC.

Priorities for the Near Future

Although much progress has been made during the last decade, there is still a long way to go before quality-of-life evaluation can be regarded as an integrated part of standard cancer clinical practice. The rapid growth in the number of EORTC studies that include quality of life as an end point may reflect the increasing awareness and importance of the subject on the part of the investigators, but it has also pointed out more clearly the flaws and shortcomings in this new field of research. The EORTC has set the following priorities for its activities related to quality-of-life issues:

Good-Quality Studies

Since EORTC trials are conducted in an international, multicenter setting, it is extremely important to have a good infrastructure and a standard approach to the collection and analysis of quality-of-life data. To ensure adequate rates of patient accrual, compliance, and data quality, there is an urgent need for a number of standard data management strategies. These strategies include implementation procedures, detailed instructions for data collection, explicit instructions on the administration of quality-of-life instruments, regulations on coding of data, and interpretation of missing data and incomplete forms. A standard training course for people who are responsible for data collection will be developed and conducted at regular intervals in all countries that participate in EORTC studies to ensure optimal benefit.

Analysis and Interpretation of Data

Despite the research efforts of the last two decades, a number of questions with regard to specific issues remain open. An important fact is that there is no optimal method for analyzing quality-of-life data. Several methods can be used and perhaps should be used to provide insight into the data. However, each method has its advantages and disadvantages, and different models have different assumptions that are not always met.

The interpretation of results is impeded by the lack of standards concerning what can be considered as a clinically important change in any quality-of-life score and the absence of standard methods to define effect sizes and to calculate sample size requirements. An important step forward would be the availability of large datasets that could be utilized in future trials for the computation of expected differences and sample sizes. Since the EORTC QLQ-C30 is currently being used in many studies, reliable datasets should become available in the near future.

A final methodologic issue relates to the integration of different outcome measures. As stated previously, cancer clinical trials have a history of parameters, all related to length-of-life outcomes. Further development of methods to combine length-of-life with quality-of-life data is both warranted and a major challenge. Since resources for health expenditure are becoming more restricted, health economic issues have become increasingly important also in cancer clinical trials. Combining economic data with quality-of-life and length-of-life data, therefore, will become increasingly important. These issues will be addressed in close collaboration with the EORTC Health Economics Unit.

Theoretical Issues

Although it has become virtually impossible nowadays to keep up with the stream of publications of empirical studies on quality-of-life issues, the theoretical foundation and framework on quality of life are still rather weak. Quality of life is a dynamic concept, like illness. However, the way in which and degree to which these two concepts interact with each other and what other additional factors may have an influence are still largely unknown. One such additional factor is the unknown role culture plays in quality-of-life issues. A unique characteristic of the EORTC is that its clinical trials are by definition cross-national studies. The total number of countries that are involved in EORTC studies is at present 31. This feature provides a treasure of information to investigate cross-cultural differences. So far, this investigation has not been done, but cross-cultural differences will become one of the major new research questions in the near future.

In conclusion, this article has outlined the experience and perspective of quality-of-life research within the EORTC. Although much progress has been made, there is still a lot of work to do before quality of life achieves its rightful place in cancer therapy evaluation. With the present enthusiasm and motivation on the part of all parties involved, we are optimistic that this process will lead to a better understanding of the impact of anticancer therapy on patients' quality of life.

References

(*1*) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.

(*2*) Aaronson NK, Cull A, Kaasa S, Sprangers MA. The EORTC modular approach to quality-of-life assessment in oncology. Int J Mental Health 1994;23:75-96.

(*3*) Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M. The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. Eur J Cancer 1994;30:635-42.

(*4*) Bjordal K, Ahlner-Elmqvist M, Tollesson E, Jensen AB, Razavi D, Maher EJ, et al. Development of a European Organization for Research and Treatment of Cancer (EORTC) questionnaire module to be used in quality of life assessments in head and neck patients. EORTC Quality of Life Study Group. Acta Oncol 1994;33:879-85.

(*5*) Presant CA, Quality of life in cancer patients. Who measures what? Am J Clin Oncol 1984;7:571-3.

(*6*) Olver JN, Matthews JP, Bishop JF, Smith RA. The roles of patient and observer assessment in anti-emetic trials. Eur J Cancer 1994;30A:1223-7.

(*7*) Aaronson NK, Ahmedzai S, Bullinger M, et al. The EORTC core quality-of-life questionnaire: interim results of an international field study. In: Osoba D, editor. Effect of cancer on quality of life. Boca Raton (FL): CRC Press, 1991:185-203.

Assessment of Quality of Life in Clinical Trials of the British Medical Research Council

*David Machin**

This article describes aspects of the way the Cancer Therapy Committee of the British Medical Research Council incorporates quality-of-life (QOL) assessments in randomized clinical trials in patients with cancer. The steps taken in incorporating QOL assessments in individual trial protocols are described. The aspects described concern problems associated with choice of instruments, time of assessment, sample size, and analysis. A protocol for patients with small-cell lung cancer that compares oral etoposide with intravenous multidrug chemotherapy is used for illustration. [Monogr Natl Cancer Inst 1996;20:97-102]

Cancer clinical trials sponsored by the British Medical Research Council (MRC) are usually organized under the auspices of either the Leukemia Working Party or the Cancer Therapy Committee. The latter has site-specific working parties who address therapeutic questions regarding solid tumors, and it is the work of the Cancer Therapy Committee that is described here.

The Cancer Therapy Committee essentially is made up of the chairs of the site-specific working parties, an independent chair, several other independent assessors, including one with special interest in quality of life (QOL), and the chief medical statistician of the MRC Cancer Trials Office. Proposals for clinical trials are generated within the site-specific working parties, and a brief summary of these proposals is presented to the Cancer Therapy Committee for its approval. Approval is or is not given at a full meeting of the committee. The particular trial coordinator of the proposal under discussion attends this meeting to explain the rationale for the trial. The coordinator is absent when a decision on the particular protocol is made. Approval of the protocol at this stage guarantees the statistical support of the Cancer Trials Office and signals the development of a full protocol. This protocol, together with the appropriate data forms, is then subsequently put to an independent Protocol Review Committee for approval. The Protocol Review Committee discusses the protocol line by line with the trial coordinator, statistician, and data manager assigned to that particular protocol. At this review, it is not usually expected that major changes will be made to the therapeutic questions being addressed, as these have been examined in detail at the earlier stages. Rather, the review is to see that the protocol is indeed practicable. Once the Protocol Review Committee gives its approval, the final protocol documentation is prepared and the trial is launched at a convenient date. Ethical approval of each protocol is given at a local level, usually by a committee of the institute where the participating clinicians work. If a trial proposed is a pragmatic one, which may require many thousands of patients, then this trial is usually coordinated through the U.K. Coordinating Committee for Cancer Research, and this type of trial is not considered further here. [*See*, for example, details of the AXIS trial, 1994 (*1*).]

Until recently, each working party had the responsibility to decide whether or not QOL was appropriate for the particular study in question; again, until relatively recently, the major use of QOL measures was confined to the Lung Cancer Working Party. This particular group has a long history of using QOL measures and was responsible for developing the MRC patient diary card (*2-4*). As will be illustrated below, this working party has made extensive use of the Rotterdam Symptom Checklist (*5*) and the Hospital Anxiety and Depression Scale (*6*) and has recently started to use the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30 (*7*).

In 1993, the MRC reached a concordat with the U.K. Department of Health. One of the consequences of that concordat was that QOL (and health economic assessment) ought to be an integral part of clinical trials. Thus, the site-specific working parties of the Cancer Therapy Committee now have to state why QOL should not be included in a particular trial. An example of a case in which QOL is not included is a trial in operable osteosarcoma. Patients with operable osteosarcoma are mainly children, and no validated QOL instrument for children was available prior to the launch of the trial in 1993. In that case, the reason for not conducting QOL was at a very practical level. Work on an appropriate instrument is in progress.

Of the 25 open, randomized phase III trials of the Cancer Therapy Committee, 12 involved QOL assessments of one form or another. QOL assessments were included in trials of cancers of the bladder, brain, colorectum, lung, prostate, kidney, and stomach. The specific reasons for inclusion of QOL assessment in a renal trial were documented (*8*).

In explanatory trials in which it is anticipated that the therapy being tested may bring more than modest therapeutic gain, as expressed in terms of patient survival, survival is used as the main outcome measure, and patient numbers are calculated on the basis of the anticipated survival benefit. In contrast, however, in some of the palliative trials of the Lung Cancer Working Party, in which attempts have been made to reduce therapy as compared with standard therapy, these trials aim for survival

**Correspondence to:* David Machin, M.D., Medical Research Council., Cancer Trials Office, 5 Shaftesbury Rd., Cambridge CB2 2BW, U.K.

equivalence. As a consequence, the QOL issues become more prominent and indeed may be the major outcome variable. If this is the case, then QOL becomes the focus for the design and, in particular, determines the end points for calculations of patient numbers.

Within the MRC and elsewhere, there is considerable experience in estimating appropriate sample sizes on the basis of survival end points (*9,10*). This is not only because such calculations have been used frequently by the statisticians but also because the clinicians are able to balance the anticipated survival gain against the weight of therapy in the patient groups and thereby determine a clinically worthwhile difference to be established by the trial. On the other hand, although QOL is clearly an important end point for the patient, experience of assessment and perhaps more importantly the "feel" for what constitutes an improvement in QOL are more problematic. Strategies that attempt to summarize subjective clinical opinion at the design stage do not appear to have been utilized (*11*).

Thus, objective definitions of what constitutes a clinically important benefit in terms of QOL have not been identified, although work is in progress in this area (*12*). To date, this problem has been circumvented somewhat for the purposes of sample size calculation by focusing on a single item or component of the QOL questionnaire and using this as a surrogate. Thus, for example, the 10 symptoms that most trouble patients with lung cancer have been identified, and the palliation of a prespecified number of these symptoms has been regarded as an indicator of a clinically important QOL improvement (*13*).

Clearly, other issues related to QOL have to be addressed. These issues include patient compliance, patient attrition, and missing data. All of these issues need to be considered at the design stage and may influence the number of patients to be recruited.

An Example

To illustrate the various aspects of development of a protocol involving QOL as an integral part, we use a randomized, controlled clinical trial of oral etoposide versus intravenous multidrug chemotherapy for the palliative treatment of patients with small-cell lung cancer and a poor prognosis. This trial, referred to as LU16, was begun in August 1992, and it is anticipated that it will close toward the end of 1996 after 500 patients are recruited.

Design

The trial design of the LU16 trial is shown in Fig. 1. The figure summarizes the eligible patients and the randomization to oral etoposide against the intravenous multidrug chemotherapy EV (i.e., etoposide + vincristine) or CAV (i.e., cyclophosphamide + doxorubicin + vincristine) and details the follow-up for QOL assessments by means of the Rotterdam Symptom Checklist and the Hospital Anxiety and Depression Scale and the period during which the patient diary card should be completed. The patient diary card was specifically included here in order to assess the influence of active therapy on those aspects of QOL that may be transitory during the treatment phase and caused either by an immediate benefit of therapy or as a consequence of the side effects of the therapy. For example, the use of the patient diary card in a previous trial had indicated transient dysphagia between 10 and 21 days from the start of radiotherapy in patients with non-small-cell lung cancer receiving a two-fraction course of radiotherapy, while such an excess was not noted for those patients randomly assigned to receive a single-fraction regimen (*14*).

The QOL assessments by means of the Rotterdam Symptom Checklist, the Hospital Anxiety and Depression Scale, and the patient diary card are completed following the schedule summarized in Fig. 1. Thus, immediately before the therapy is started and before the randomized treatment is allocated, the patient completes each of these three instruments. The daily diary card is then completed for 12 weeks, which covers the period until completion of the fourth cycle of chemotherapy. The Rotterdam Symptom Checklist and the Hospital Anxiety and Depression Scale are completed every 3 weeks immediately before the chemotherapy is administered and thereafter until 3 months, then monthly to 6 months, then every 2 months to 1 year, and every 3 months thereafter.

Since the treatment was scheduled to be completed by the 12th week (3 months), this was believed to be not only an appropriate date for QOL assessment but also the key QOL assessment for evaluation and hence design purposes (Fig. 2).

In many situations, it is not always clear when, for example, the Rotterdam Symptom Checklist or the Hospital Anxiety and Depression Scale questionnaire should be completed in order to make sensible comparisons between treatments, especially if the alternative therapies under test are of different types (e.g., chemotherapy as opposed to radiotherapy) and/or of different duration. It is usually not desirable to ask for additional clinic visits to complete QOL instruments alone merely in order to maintain synchrony between treatment assessments. Usually some compromise has to be reached between the optimal time points that are best for the scientific question posed and the demands of everyday patient care.

Number of Patients

The stated objectives of LU16 trial are listed in Fig. 2, which identifies palliation as the major outcome variable. As already referred to, there is an intrinsic difficulty in defining benefit in these circumstances. To assess patient numbers, it was thought appropriate to select a series of symptoms from the Rotterdam Symptom Checklist to form the basis of a definition of palliation (*13*). The symptoms identified were cough, pain, anorexia, and shortness of breath. Each of these symptoms is scored on an ordered categorical scale from 0 (not at all) to 3 (very much). Thus, the score at presentation could range from 0 to 12. With appropriately selected patients, however, the lower limit is unlikely to be less than 2. This definition, albeit somewhat arbitrary, was then applied to patient data from previous trials and was found to achieve approximately 50% palliation (improvement in QOL) in the equivalent of the CAV arm.

These calculations led to a sample size of 400 patients, but because of patient attrition it was believed to be appropriate to increase this sample size (*15*). A judgment was then made suggesting that 500 patients would be more appropriate. The corresponding statement made in the protocol is shown in Fig. 3. It

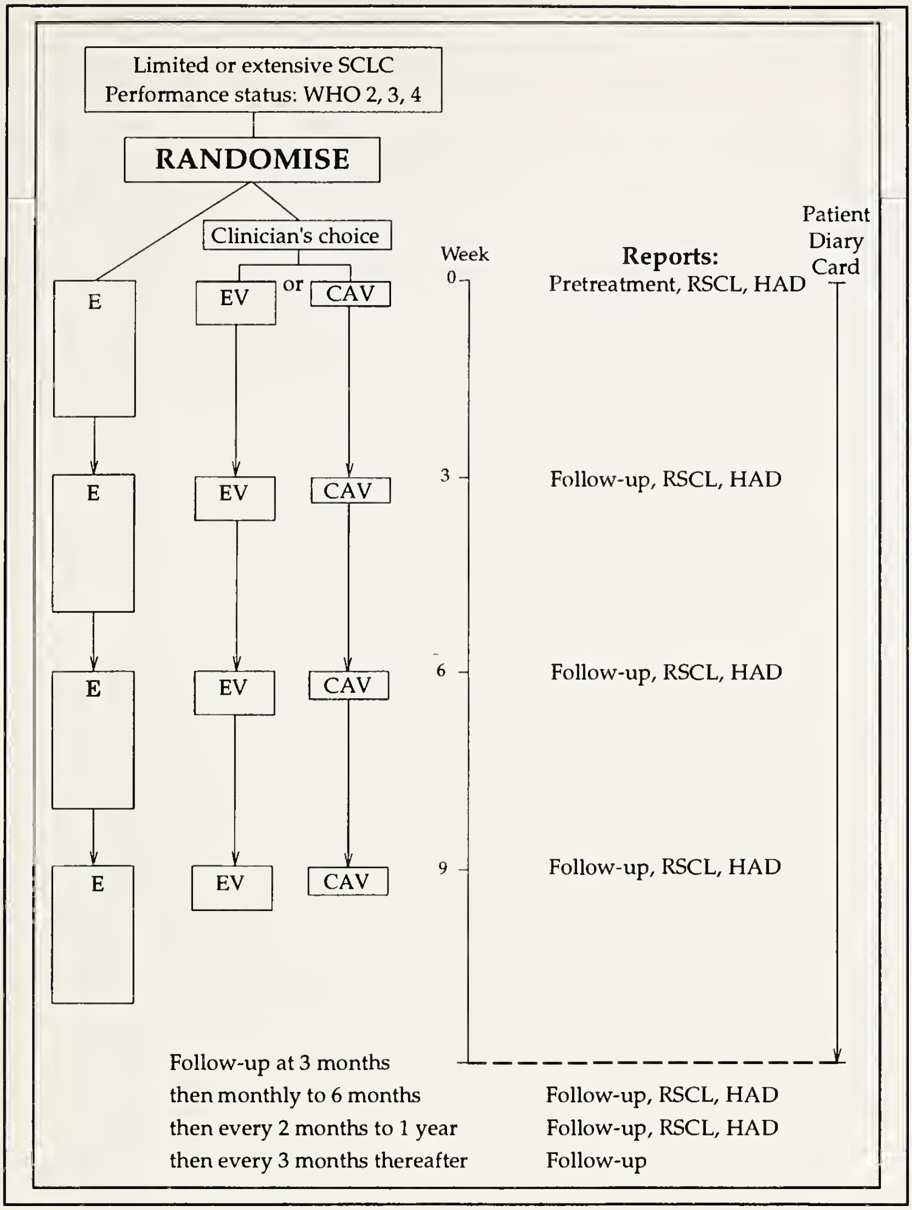

Fig. 1. LU16 trial design: British Medical Research Council randomized, controlled clinical trial of oral etoposide versus intravenous multidrug chemotherapy in the palliative treatment of patients with small-cell lung cancer (SCLC) and poor prognosis (dated August 1992). The panels that are presented in this section of the article are extracted from the LU16 protocol itself and have not been edited. E = oral etoposide; EV = etoposide + vincristine; CAV = cyclophosphamide + doxorubicin + vincristine; RSCL = Rotterdam Symptom Checklist; HADS = Hospital Anxiety and Depression Scale; WHO = World Health Organization.

was recognized that comparisons between treatments with respect to other aspects of QOL would be made on a more informal (exploratory) basis.

Comparison of QOL Instruments

Since the launch of the above trial, the copyright format of the EORTC core questionnaire (EORTC QLQ-C30) has become available. There is therefore some debate as to whether or not this format should replace the Rotterdam Symptom Checklist. As a consequence, rather than all patients on the LU16 trial receiving the Rotterdam Symptom Checklist as is indicated by Fig. 1, half of the patients now receive this checklist and half receive the EORTC QLQ-C30 on a random basis. This randomization gives the trial a 2×2 factorial design format, although analysis of the two instruments cannot be made on this basis. In a certain sense, the best comparison of the two instruments should be a within-patient comparison, with both instruments completed almost simultaneously, albeit this design has obvious flaws. In any event, it is recognized that this is not possible, at least within the context of a randomized, controlled trial involving many centers, as it clearly places an extra burden on both patients and staff. Since the primary objective of a trial is to compare treatments, it will be of interest to see which instrument best reflects the (standardized) true treatment difference. There is some circularity here, since we do not know the true treatment difference (*16*). Such a comparison is also likely to involve other factors in the final choice of instrument for future use. These factors include, in particular, considerations of any major differential in compliance rates.

Practical Considerations

One of the major obstacles to recruitment to clinical trials is often the complexity of the trials themselves in terms of the extra information on a patient that it is necessary to record over

PRINCIPAL ENDPOINT:

1. Palliation of major symptoms at 3 months

SECONDARY ENDPOINTS:

2. Adverse effects of treatment
3. Quality of life
4. Survival
5. Response

- Palliation of major symptoms
- Palliation is defined as having a reduction in the sum of the cough, pain, anorexia and shortness of breath scores at three months from randomisation
- Patients who die before 3 months (whether or not they have palliation) are defined as failures of palliation

Fig. 2. End points and definition of principal end point for the LU16 protocol (dated August 1992).

It is anticipated that major symptoms (cough, pain, anorexia and shortness of breath) will be palliated in 50% in the control group within the first 3 months of treatment. The oral etoposide treatment will be regarded as equivalent to the intravenous chemotherapy treatment if palliation is achieved in not less than 37.5% of the patients. With this 12.5% level of equivalence, a one-sided test at 5% and 80% power would require a total of between 400 and 500 patients.

Fig. 3. Statistical considerations section of the LU16 protocol (dated August 1992).

and above that recorded in routine clinical practice. Of course, there are other more difficult areas, including seeking informed consent from the patients (*17*). As a consequence of the recognized burden on the clinical team, a great emphasis, at least in Europe, has been on conducting minimum-forms trials. The clinicians themselves have recognized and welcomed the need for such an approach. It is therefore somewhat counter to this trend that we now add (many) QOL assessments. Of course, these assessments are intended to be completed by the patients themselves. At the very least, however, the participating centers need to make appropriate arrangements for distribution, completion, and return to the trials office. This is no small task. Suggestions to individual centers as to how they may facilitate completion of the QOL instruments are usually included in the study protocol, and some advice for the LU16 trial is summarized in Fig. 4.

The burden on the trials offices themselves is also not easy to dismiss. The data received have to be processed, their quality needs to be assessed and queried, missing forms have to be pursued, and finally analysis needs to be conducted. Thus, QOL assessments, albeit a desirable feature for the majority of randomized trials of treatments for cancer patients, should not be conducted without taking account of the resources required.

Analysis

Although it is not the purpose of this article to go into details of aspects of analysis of QOL data once collected, this is clearly an important issue, and steps that have been taken by the MRC in this respect have been outlined elsewhere (*18*). These steps follow lines similar to those suggested for the analysis of menstrual bleeding diaries (*19*). The patient-diary-card assessments have been reported both in terms of relative compliance between treatments and by means of a daily summary measure of individual symptoms (*20*).

For example, Fig. 5 shows the patient-diary-card profile as recorded for activity in a randomized trial comparing ECMV chemotherapy (i.e., etoposide + cyclophosphamide + metho-

Application of the Quality of Life Questionnaires

It is important to explain to the patient that the Rotterdam Symptom Checklist (RSCL) and the Hospital Anxiety and Depression (HAD) scale refer to how they have been feeling during the past week, and that all questions should be answered even if the patient feels them to be irrelevant. Emphasise that the completion of these forms helps doctors find out more about the effects of the treatment. Also remind the patient to complete the back of the RSCL. The patient should complete the questionnaires, without conferring, whilst waiting to be seen in the clinic. Collect the questionnaires before the patient leaves and check that all questions have been answered, if necessary going back to the patient immediately and asking them to complete any missing items.

Fig. 4. Application of the quality-of-life questionnaires in the LU16 trial (dated August 1992).

trexate + vincristine) with selective palliative treatment in 162 patients with small-cell lung cancer. This profile was not included in the published report (*21*). Activity was recorded on a 5-point scale, ranging from 1 (at work or active retirement) to 4 (confined to home or hospital) to 5 (confined to bed). Thus, Fig. 5 indicates that the ECMV treatment, which requires hospitalization, does indeed induce more inactivity than the selective therapy in the first few days following randomization. Thereafter, activity levels improve and are comparable between the two treatment modalities. No formal testing of such profiles is attempted. The compliance for each treatment on a monthly basis is indicated beneath the horizontal axis in Fig. 5 and clearly indicates how poor it was, although this trial was conducted between 1981 and 1985 and organizational details for encouraging completion of patient diary cards have since been improved (Fig. 4).

There are certain hidden problems with this method of summary, however. These problems include patient attrition and the

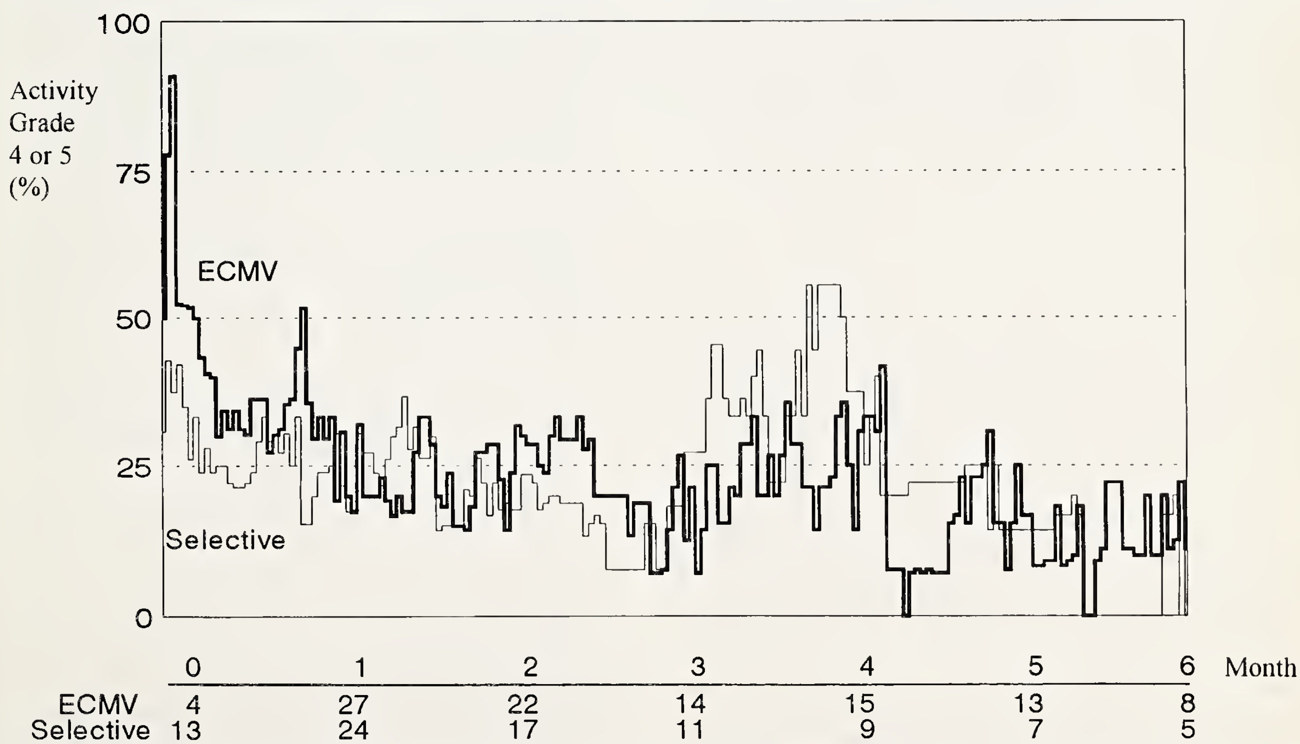

Fig. 5. Patient diary profiles of patients randomly assigned to receive either ECMV (i.e., etoposide + cyclophosphamide + methotrexate + vincristine) or selective treatment with respect to activity as assessed by a patient diary card.

"blur" that occurs if, for example, patients receive their treatment on days other than those scheduled. Thus, if the patient diary card was assessing nausea or vomiting during an intensive chemotherapy regimen, this symptom would be greatest on treatment days but more or less absent on other days. Such a profile would be represented by spikes at the appropriate cycle day interspersed by a very low plateau if all therapy was on schedule but blurred otherwise.

For other QOL instruments that are not recorded on a daily basis, it is therefore important that these analyses in a sense compare like with like. Thus, one strategy adopted is to make treatment comparisons between patients completing the same number of QOL questionnaires (and at the same time points) and then to combine these differences by means of a stratified analysis as one might do in any standard survival-type comparison. The summary statistics used in such comparisons have usually been the slope and intercept of a linear regression equation fitted to the individual patient profiles. These summary statistics are then summed over patients, and treatment comparisons are made with these. This method can be extended to include orthogonal polynomial fits if changes over time are not even approximately linear.

In this respect, we prefer the approach to repeated measures data advocated by Matthews et al. (*22*), who suggested that key features of each profile be identified, such as the area under the curve (AUC), rather than a formal repeated analysis of variance that can be utilized through standard statistical packages. The main reason for our preference is that it is important that any analysis focuses on aspects of QOL summary that have a relatively easy interpretation. It is recognized, however, that summarizing such complex data by means of relatively few parameters may hide more subtle treatment differences that nevertheless may have an important impact on a patient's well-being. Approaches to analysis suggested by Korn (*23*) also concerned the AUC, and work is in progress to confirm the utility of this particular approach.

Discussion

The introduction of QOL assessments into the conduct of randomized clinical trials in cancer raises issues that range from the choice (and perhaps development) of an appropriate instrument, choice of completion times, additional burden on the patient and the trial itself, appropriate sample size, analysis, and interpretation. Of particular importance here is any "trade off" between QOL and survival. Increased survival should not necessarily dominate, particularly if it is at some considerable cost in terms of QOL, but neither should the opposite be seen to be the case. An appropriate estimate of survival and an equally reliable quantification of QOL (or at least aspects thereof) are likely to be jointly valuable guides to patient management. The role of QOL in other areas of MRC activities has been described in part by Johnson (*24*).

References

(*1*) The AXIS colorectal cancer trial: randomisation of over 2000 patients. The AXIS Steering Group. Br J Surg 1994;81:1672.

(*2*) Fayers PM, Jones DR, Girling DJ. Measurement of quality of life in cancer clinical trials. Cancer Treat Sympos 1985;2:25-30.

(*3*) Jones DR, Fayers PM, Simmons J. Measuring and analyzing quality of life in cancer clinical trials. In: Aaronson NK, Beckman J, editors. The quality of life of cancer patients. New York: Raven Press, 1987:41-61.

(*4*) Fayers PM, Bleehen NM, Girling DJ, Stephens RJ. Assessment of quality of life in small-cell lung cancer using a Daily Diary Card developed by the Medical Research Council Lung Cancer Working Party. Br J Cancer 1991;64:299-306.

(*5*) De Haes JC, van Knippenberg FC, Neijt JP. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. Br J Cancer 1990;62:1034-8.

(*6*) Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361-70.

(*7*) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.

(*8*) Fayers PM, Cook PA, Machin D, Donaldson N, Whitehead J, Ritchie A, et al. On the development of the Medical Research Council trial of alpha-interferon in metastatic renal carcinoma. Urological Working Party Renal Carcinoma Subgroup. Stat Med 1994;13:2249-60.

(*9*) Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat Med 1982;1:121-9.

(*10*) Fayers PM, Machin D. How many patients are necessary? Br J Cancer 1995;72:1-9.

(*11*) Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. Stat Med 1986;5:1-13.

(*12*) George S, Julious SA, Campbell MJ. Sample sizes for studies using SF-36. J Epidemiol Commun Health. In press.

(*13*) Hopwood P, Stephens RJ. Symptoms at presentation for treatment in patients with lung cancer: implications for the evaluation of palliative treatment. The Medical Research Council (MRC) Lung Cancer Working Party. Br J Cancer 1995;71:633-6.

(*14*) A Medical Research Council (MRC) randomised trial of palliative radiotherapy with two fractions or a single fraction in patients with inoperable non-small-cell lung cancer (NSCLC) and poor performance status. Medical Research Council Lung Cancer Working Party. Br J Cancer 1992;65:934-41.

(*15*) Machin D, Campbell MJ. Statistical tables for the design of clinical trials. Oxford: Blackwell Scientific Publications, 1987:94-131.

(*16*) Machin D, Lewith GL, Wylson S. Pain measurement in randomised clinical trials: a comparison of two pain scales. Clin J Pain 1988;4:161-8.

(*17*) Altman DG, Whitehead J, Parmar MK, Stenning SP, Fayers PM, Machin D. Randomised consent designs in cancer clinical trials. Eur J Cancer. In press.

(*18*) Hopwood P, Stephens RJ, Machin D. Approaches to the analysis of quality of life data: experiences gained from a Medical Research Council Lung Cancer Working Party palliative chemotherapy trial. Qual Life Res 1994;3:339-52.

(*19*) Machin D, Farley TM, Busca B, Campbell MJ, d'Arcangues C. Assessing changes in vaginal bleeding patterns in contracepting women. Contraception 1988;38:165-79.

(*20*) Bleehen NM, Girling DJ, Machin D, Stephens RJ. A randomised trial of three or six courses of etoposide, cyclophosphamide, methotrexate, and vincristine or six courses of etoposide and ifosfamide in small cell lung cancer (SCLC). II: Quality of life. Medical Research Council Lung Cancer Working Party. Br J Cancer 1993;68:1157-66.

(*21*) Survival, adverse reactions and quality of life during combination chemotherapy compared with selective palliative treatment for small-cell lung cancer. Report to the Medical Research Council by Its Lung Cancer Working Party. Respir Med 1989;83:51-8.

(*22*) Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research [*see* comment citation in Medline]. BMJ 1990;300:230-5.

(*23*) Korn EL. On estimating the distribution function for quality of life in cancer clinical trials. Biometrika 1993;80:535-42.

(*24*) Johnson AL. Some statistical issues in quality of life measurements. In: Trimble MR, Dodson WE, editors. Epilepsy and quality of life. New York: Raven Press 1994;65-84.

United Kingdom Cancer Research Campaign Approach to Quality-of-Life Research in Cancer Clinical Trials

*Penelope Hopwood**

Clinical trials of new anticancer therapies form an important part of the research activity of the Cancer Research Campaign (United Kingdom), and quality-of-life (QOL) end points are being increasingly used in the evaluation of new treatment approaches. The Campaign has a unique policy of supporting a broad range of scientific and clinical research, including psychosocial studies, and thus QOL research is generated in a variety of clinical settings. The focus of interest for the Cancer Research Campaign lies in QOL design and assessment rather than the routine application of QOL protocols. Clinical investigators are free to adopt an individual approach, but the Campaign operates a strict peer-review system in protocol assessment. Some standardization of approach is being achieved through consensus of opinion and wide collaboration, both nationally and internationally. [Monogr Natl Cancer Inst 1996;20:103-5]

Table 1. Cancer Research Campaign

Scientific and clinical research	Educational and psychosocial research (EPR)
Scientific committee	EPR committee
↓	↓
Funding of clinical and scientific research, via	Funding of psychosocial and QOL research, via
Clinical trials centers	Research groups
Research groups	Program grants
Program grants	Project grants
Project grants	Fellowships
Fellowships	Studentships
Studentships	Collaborative research initiatives
Collaborative research initiatives	
Clinical trials	
QOL protocols	

The Cancer Research Campaign (CRC) (a United Kingdom national charity) supports a wide-ranging portfolio of research encompassing the nature and causes of cancer, new approaches to treatment and prevention, clinical trials of new therapies, psychosocial studies, and an educational program. Clinical and nonclinical training programs are also funded and a number of personal fellowships are awarded.

The CRC is unique in the funding of cancer research in the United Kingdom because of its policy of supporting a broad range of scientific and clinical activities. This, in turn, means that quality-of-life (QOL) research can originate from many different clinical and/or academic sites, which are shown in Table 1.

The CRC is rigorous in the application of peer review in the assessment of research, using the expertise of its own committees and external, often international, referees. Therefore, when directly funded by the CRC, QOL protocols are also subject to this close scrutiny, which ensures that a high standard is achieved.

The funding of the educational and psychosocial research program, which includes a small number of QOL projects, accounts for approximately 5% of the overall budget and is assessed and administered through a separate committee. While QOL protocols are more likely to arise within the clinical trials setting, some specific projects, such as the QOL study in the U.K. Tamoxifen Chemoprevention Trial, have been funded directly from the educational and psychosocial research budget.

Clinical trials originate from individuals or groups within universities, hospitals, and medical schools and also through project grants. The major trial centers now incorporate QOL end points in most trials, principally phase III randomized clinical trials, but also in some phase II studies.

The CRC hosts an expert committee of leading scientists and clinicians involved in the development of new therapy for cancer, i.e., the Phase I/II Clinical Trials Committee. This highly qualified committee advises on the development and testing of novel anticancer agents and carries out early (phase I and II) clinical trials. The committee's policy, like that of the European Organization for Research and Treatment of Cancer (EORTC), is not to conduct QOL research in these early stages of clinical testing of new drugs. A wide range of cancers is covered by CRC phase III trials, including all major solid tumor sites, lymphomas, and hematologic cancers.

One particular focus of activity is the CRC Cancer Trials Office, located at Kings College Hospital, which supports the CRC Breast Cancer Trials Group. The group structure comprises four working parties (responsible for biological protocols, new

**Correspondence to:* Penelope Hopwood, M.D., Cancer Research Campaign, Psychological Medicine Group, Christie Hospital NHS Trust, Stanley House, Wilmslow Rd., Withington, Manchester M20 4BX, United Kingdom.

See "Note" section following "References."

studies, current adjuvant trials, and closed trials) that interact with the central trials group parent committee. An executive committee, which includes an expert on QOL (L. Fallowfield), coordinates the activity of the different subgroups. In this way, a consistent approach to QOL research is ensured, and the structure facilitates the design, development, and analysis of QOL end points in a cohesive way.

Several major trial centers (e.g., Birmingham and Glasgow) have recently appointed a person to be responsible for QOL research so that this can be developed and coordinated efficiently.

In addition, CRC clinicians are involved in collaborative research with the U.S. National Cancer Institute, the EORTC, the British Medical Research Council, and the Imperial Cancer Research Fund. The United Kingdom Coordinating Committee for Cancer Research (UKCCCR) acts as a coordinating committee for cancer research in many of the U.K. collaborative programs, which are also starting to include QOL protocols. An example is the UKCCCR Adjuvant Breast Cancer Trial.

QOL Application

While QOL research is primarily associated with the cancer trials that are described above, its application is much wider. QOL measures have been incorporated in CRC-funded research evaluating psychosocial interventions in controlled randomized trials (*1*) and are currently being used in psychosocial studies of women with a genetically high risk of cancer.

In the field of cancer prevention, a battery of self-report questionnaires is being administered to women in a randomized trial of tamoxifen versus placebo. Psychosocial researchers funded by the Campaign incorporate QOL measures in a wide range of projects. This adds to the overall expertise in generating and analyzing QOL data, to the development of subscales, and to the refinement of measures for use in the clinical trial setting.

Incorporating QOL Into Phase III Clinical Trial Protocols

To date, the CRC has not published a mission statement advocating routine incorporation of QOL into phase III clinical trials, in contrast to the policy advocated, for example, by the National Cancer Institute of Canada, although the British Medical Research Council now expects to see QOL assessments in trial protocols. Nevertheless, U.K. investigators in CRC clinical trials are being asked increasingly, by protocol review committees and peer group referees, to consider adding QOL end points where appropriate alongside the more traditional outcome measures. There is also growing interest from purchasers and providers in generating these data. Consequently, there is evidence that the integration of QOL research in clinical trials has expanded considerably over recent years, and through informal collaboration and the open exchange of ideas, a considerable degree of overlap in approach has developed.

In designing QOL protocols, there is agreement among QOL researchers that such studies should answer a specific research question and (where evidence exists from earlier research) should test a hypothesis. Table 2 shows elements of the decision process that may be considered when assessing the potential inclusion of QOL end points. This is more likely to lead to a proper consideration of the sample size, selection of appropriate measures, timing of assessment, duration of research, and other issues of methodology. It is also more likely to ensure that the results have clinical relevance and practical use. It is important that QOL end points are not included without this kind of planning, since the research makes considerable demands on resources and must be justified. Also, poorly planned research is more likely to generate incomplete data, precluding any useful interpretation of the results. Thus, wherever possible, QOL protocols should be designed in parallel with the clinical trial and in collaboration with someone with specialist knowledge of QOL.

Table 2. Deciding when to assess QOL in clinical trials: a decision tree

Is there likely to be a difference in the treatments compared that will have an impact on QOL?	→What is the principal QOL research question?
Has the impact on QOL been measured before?	→Does it warrant replication?
Is the expected effect of treatment easily measurable?	→Should a pilot study be conducted? Is there a suitable instrument?
Is the sample big enough to detect a difference?	→Should collaboration be considered?
Is the potential workload/cost to the patient/staff/institution acceptable?	→Will the results of the QOL study influence future patient care?

The area of QOL design and assessment is of particular interest to the CRC and one where it is most willing to provide research funding rather than the routine application of QOL protocols.

QOL Measures

Investigators are free to select the most appropriate measure(s) for any particular study; there is no overall policy to limit this. However, a number of groups may have been influenced by published recommendations made by a working party of the Medical Research Council Cancer Therapy Committee (*2*) that suggested that the Rotterdam Symptom Checklist (*3*) combined with the Hospital Anxiety and Depression Scale (*4*) provided the optimal approach at the time. The recommendations were made in an effort to encourage some degree of commonality of measures in trials, to ensure compatibility of results. Since that time, a number of other carefully developed and well-validated measures have been published: for example, the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) (*5*) and the Functional Assessment of Cancer Therapy—General Scale (FACT-G) (*6*). These, and a limited number of other instruments, are all currently being used in cancer trials. More specific measures of body image, sexual adjustment, and attitudes to illness are being developed by CRC research fellows.

In summary, the CRC is actively involved in QOL research in cancer clinical trials and, as a result of its broad support of psychosocial research, is also able to support the necessary

developmental and advisory functions through researchers in the psychosocial field. Active collaboration with other organizations and institutions ensures that some degree of standardization and cross-fertilization of ideas is achieved and facilitates collaboration in large multicenter and, in some cases, multinational trials.

References

(*1*) Greer S, Moorey S, Baruch JD, Watson M, Robertson BM, Mason A, et al. Adjuvant psychological therapy for patients with cancer: a prospective randomised trial [*see* comment citations in Medline]. BMJ 1992;304:675-80.

(*2*) Maguire P, Selby P. Assessing quality of life in cancer patients. Br J Cancer 1989;60:437-40.

(*3*) de Haes JC, van Knippenberg FC, Neijt JP. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. Br J Cancer 1990;62:1034-8.

(*4*) Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361-70.

(*5*) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.

(*6*) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:572-9.

Note

Supported by the Cancer Research Campaign, United Kingdom.

Health-Related Quality-of-Life Studies of the National Cancer Institute of Canada Clinical Trials Group

*David Osoba, Janet Dancey, Benny Zee, James Myles, Joseph Pater**

Since 1989, the National Cancer Institute of Canada Clinical Trials Group (NCIC CTG) has been successful in implementing and completing health-related quality-of-life (HQL) assessments as part of phase III clinical trials. Compliance rates for completing HQL instruments remain high, with a minimal amount of missing data. It is believed that this success is attributable not only to the high degree of commitment to measuring HQL by clinical trials investigators, nurses, data managers, and central office administrative staff, but also to the educational process that was instituted after the development of a CTG policy for measuring HQL. From inception to May 1995, a total of 27 clinical trials with HQL assessment have been initiated or completed. In the majority of trials, the core HQL instrument is the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30). In addition to answering specific questions about HQL in these clinical trials, the trials provide the opportunity to do research into the measurement of HQL. Thus, current clinical trials include research questions about the appropriate timing of assessments, the reliability and validity of the QLQ-C30 and other instruments, the role of HQL data in assessing toxicity, and the significance of the results of HQL assessments. It is anticipated that this activity not only will be a rich source of information about the effects of cancer and its treatment on HQL but also will lead to improvements in measuring HQL in oncology. [Monogr Natl Cancer Inst 1996;20:107-11]

The National Cancer Institute of Canada Clinical Trials Group (NCIC CTG) formed a Quality-of-Life Committee in 1987 and adopted a policy for measuring health-related quality of life (HQL) in clinical trials in 1989 (*1*). The policy is that "there should be a statement about the anticipated impact on quality of life in every phase III clinical trial and whether or not quality-of-life measures will be incorporated in the protocol." The concept underlying the policy is that HQL measurement should be an integral part of a phase III trial rather than an activity added to the trial, i.e., a companion study. The implications for the implementation of this philosophy are that the central office administrative functions required for the HQL component of a trial are assigned to the same personnel who are responsible for the entire trial, and HQL activities are integrated into their usual activities rather than through a separate administrative structure. Thus, protocol development, form production, data management, and analysis are treated as integral functions within the assigned job descriptions of the existing personnel.

The Quality-of-Life Committee, consisting of volunteers from cancer centers across the country, assumed the responsibility for assisting investigators with the integration of HQL assessments into the proposed trials by developing writing guidelines for protocols (*1*); providing educational seminars for investigators, clinical trials nurses, and data managers; and providing written instructions for collecting HQL data to be used by clinical trials personnel in the participating cancer centers. Particular attention was paid to these educational processes, since it was recognized that HQL data must be collected as completely as possible at the appropriate time points. Otherwise, the result would be missing data that would interfere with making valid conclusions. The organization and functions of the Quality-of-Life Committee, the protocol-writing guidelines, and the instructions to clinical data managers have been presented in detail previously (*1*). Therefore, the remainder of this paper will concentrate on the HQL activities of the NCIC CTG since 1991.

Clinical Trials With a Quality-of-Life (QOL) Component

Current Studies

Since the adoption of the policy for measuring HQL, a total of 27 NCIC-sponsored, phase III trials containing HQL assessments have been implemented (Table 1). They range across several anatomic sites. Five trials have been studies in symptom control and seven have involved the combined use of radiation therapy and chemotherapy or radiation therapy alone, whereas the remainder are chemotherapy trials. Brief titles of the trial

**Affiliations of authors:* D. Osoba, British Columbia Cancer Agency and University of British Columbia, Vancouver, Canada; J. Dancey, B. Zee, J. Myles, J. Pater, National Cancer Institute of Canada Clinical Trials Group, Kingston, Ontario, Canada.

Correspondence to: David Osoba, M.D., Communities Oncology Program, British Columbia Cancer Agency, 600 W. 10 Ave., Vancouver, British Columbia V5Z 4E6, Canada.

and the HQL instrument used are presented in Table 2. Five trials have been completed (two are still being analyzed), whereas three were closed because of lack of accrual. Currently, 17 trials are open and accruing patients.

The results from the completed trials are of interest. In ME.7, a trial comparing interferon gamma to levamisole in the adjuvant treatment of high-risk, surgically resected primary malignant melanoma, pretreatment global QOL predicted for subsequent on-treatment global QOL (*2*). Preliminary results from two studies of the efficacy of H_3T-antagonist antiemetics with or without dexamethasone (SC.8 and SC.9) indicated that patients who experienced postchemotherapy vomiting had lower physical, role, and social function, lower QOL, and more fatigue than did patients who did not have vomiting (Osoba D, Lee B,

Table 1. Summary of studies with HQL assessment

Disease site	No.	Status
Breast	4	1 closed, 3 open
Colorectum	3	All open
Genitourinary tract	2	Both open
Gynecologic site	4	2 closed, 2 open
Head and neck	1	Planned
Hematologic site	3	1 open, 2 planned
Lung	3	1 closed, 2 open
Melanoma	1	Closed
Sarcoma	1	Open
Symptom control	5	4 closed, 1 open
Total	27	9 closed, 15 open, 3 planned

Table 2. Studies with HQL components (April 1995)*

Disease site	Symbol	Brief title†	Instruments‡
Breast	MA.5§	CMF versus CEF in patients with positive nodes	BCQ
	MA.8‖	Vr plus doxorubicin versus doxorubicin in metastatic and recurrent disease	QLQ-C30
	MA.10‖	Dose-intensive chemotherapy for locally advanced/inflammatory cancer	QLQ-C30
	MA.11‖	Escalating FEC with G-CSF	BCQ
Gastrointestinal	CO.7‖	Adjuvant 5-FU and leucovorin versus delayed therapy after resection of liver or lung metastasis in colorectal cancer	QLQ-C30, SF-36
	CO.9‖	Adjuvant high-dose versus standard-dose levamisole + 5-FU and leucovorin in colorectal cancer	QLQ-C30
	CO.10‖	Immediate versus delayed 5-FU + leucovorin in asymptomatic advanced colorectal cancer	QLQ-C30
Genitourinary	PR.3‖	Total androgen blockade ± pelvic irradiation in localized carcinoma of the prostate	QLQ-C30, FACT-P
	PR.5¶	Short radiation fractionation schedule for localized prostate cancer	QLQ-C30
Gynecology	CX.2‖	Radiation ± cisplatin for locally advanced squamous cell cancer of the cervix	QLQ-C30
	CX.3#	Cisplatin ± etoposide and ifosfamide for carcinoma of the cervix	QLQ-C30
	OV.9**	Paclitaxel in platinum-pretreated ovarian cancer	QLI
	OV.10‖	Platinum and paclitaxel versus platinum and cyclophosphamide for advanced ovarian cancer	QLQ-C30
Head and neck	HN.1¶	Elective neck dissection in early oral cancer	QLQ-C30, SF-36
Hematology	HD.6‖	Radiotherapy or ABVD + radiotherapy versus ABVD alone for early-stage Hodgkin's disease	QLQ-C30
	LY.5¶	CHOP versus CHOP + G-CSF for intermediate and high-grade non-Hodgkin's lymphoma in the elderly	QLQ-C30
	MY.7¶	Melphalan + dexamethasone or prednisone for multiple myeloma	QLQ-C30
Lung	BR.8‖	CODE versus alternating CAV and EP in extensive-stage small-cell lung cancer	QLQ-C30
	BR.9#	Chemotherapy + surgery versus radiation therapy for stage III A non–small-cell lung cancer	QLQ-C30
	BR.10‖	Adjuvant Vr and cisplatin in resected non–small-cell lung cancer	QLQ-C30
Melanoma	ME.7††	Human interferon gamma versus levamisole as adjuvant therapy for poor prognosis malignant melanoma	QLQ-C30
Sarcoma	SR.2‖	Preoperative versus postoperative radiation therapy for soft tissue sarcoma	SF-36, TESS
Symptom control	SC.8§	Ondansetron and dexamethasone in highly emetogenic chemotherapy	QLQ-C30
	SC.9§	Granisetron ± dexamethasone in moderately emetogenic chemotherapy	QLQ-C30
	SC.10#	Clodrinate versus placebo for bone pain in metastatic cancer	QLI
	SC.11§	Dolasetron mesylate versus ondansetron ± dexamethasone for moderately emetogenic chemotherapy	QLQ-C30
	SC.12‖	Dexamethasone for prophylaxis of radiation-induced emesis	QLQ-C30

*Study-specific modules are added to the above questionnaires in all studies except those involving the BCQ and QLI.

†CMF = cyclophosphamide, methotrexate, and 5-fluorouracil (5-FU); CEF = cyclophosphamide, epirubicin, and 5-FU; Vr = vinorelbine; FEC = 5-FU, epirubicin, and cyclophosphamide; G-CSF = granulocyte colony-stimulating factor; ABVD = doxorubicin, bleomycin, vinblastine, and dacarbazine; CHOP = cyclophosphamide, doxorubicin, vincristine, and prednisone; CAV = cyclophosphamide, doxorubicin, and vincristine; and EP = etoposide and prednisone.

‡BCQ = Breast Cancer Questionnaire (*3*); QLQ-C30 = EORTC Quality of Life Questionnaire consisting of 30 items or minor variations thereof (*4,5*); FACT-P = Functional Assessment of Cancer Therapy-Prostate, a variant of FACT-General (*6*); QLI = Quality of Life Index (*7*); SF-36 = Medical Outcomes Survey—short form with 36 items (*8*); and TESS = Toronto Extremity Salvage Score-University Musculoskeletal Oncology Unit, Mount Sinai Hospital, Toronto, Canada.

§Completed accrual, analysis proceeding.

‖Open, accruing patients.

¶Planned to open in 1995.

#Closed because of lack of accrual.

**Completed accrual, analysis completed.

††Completed, analysis completed, results published (*2*).

Warr D, Kaizer I, Latreille J, Pater J: manuscript submitted for publication). Since prechemotherapy QOL scores were different in patients who vomited compared with scores in those patients who did not vomit after chemotherapy, the change between these scores and postchemotherapy QOL scores was used to determine the effect of vomiting on QOL in the week following chemotherapy. Only global QOL and fatigue were adversely affected. These results have been confirmed in a larger sample of patients enrolled in SC.8 and SC.9 (*9*). Thus, QOL assessments in studies on the efficacy of antiemetics in controlling chemotherapy-induced emesis are providing new information about the effect of pretreatment QOL status on postchemotherapy vomiting and QOL.

Questionnaire Completion Rates and Missing Data

Compliance with questionnaire completion was very high in the first three studies that were analyzed (*10*). Compliance in completed trials continues to be high (Table 3). Furthermore, the rate of missing data within questionnaires is small. These appear to be acceptable rates that will allow an analysis of almost all the potential data. The high compliance rates are attributable to the efforts that were made at the outset to educate investigators, clinical trials nurses, and data managers about the importance of avoiding missing data and to the importance placed on the collection of HQL data by the personnel at the central office of the NCIC CTG.

New Directions

The policy of including HQL assessments in as many phase III trials as possible continues, but the CTG is also using the opportunity to ask additional questions about QOL within these trials. Some examples of the questions being asked are the following:

What is the appropriate timing of the HQL assessments in particular circumstances? In two studies on the effects of antiemetics on chemotherapy-induced emesis (SC.8 and SC.9), patients were asked to complete the HQL questionnaires 1 week after the chemotherapy, in part because emesis after high-dose cisplatin may last 5-6 days and in part because the time frame of the questions in the QLQ-C30 is 1 week. However, is this an appropriate time frame for moderately emetogenic chemotherapy if most of the nausea and vomiting is over 3-4 days after the chemotherapy? A more appropriate time frame might be 3 days and patients could complete the questionnaire 3 days after chemotherapy. This design has been used in SC.11, and the results are currently being analyzed.

Can some of the domains of the QLQ-C30 be revised to increase reliability? The role function domain of the QLQ-C30 has been shown to have reliability coefficients (Cronbach's alpha), ranging from 0.52 to 0.66 (*4,5*), whereas alphas for the other domains are almost always higher than 0.70. The European Organization for Research and Treatment of Cancer (EORTC) Study Group on Quality of Life reworded the two questions pertaining to role function, and these reworded questions were included in addition to the questions with the original wording in SC.11. Cronbach's alphas for the reworded questions in 696 patients varied from 0.81 to 0.88 at three time points as compared with 0.59 to 0.67 for the original wording. Other modifications to some aspects of the QLQ-C30 are also being made as a result of our studies.

Is there convergent validity between some of the popular HQL questionnaires? Niezgoda and Pater (*11*) used a multitrait-multimethod matrix in a study of 96 patients comparing the QLQ-C30 with the Sickness Impact Profile, the McGill Pain Questionnaire, the General Health Questionnaire, and the Cancer Rehabilitation Evaluation System. They concluded that the findings supported the validity of many domains of the QLQ-C30.

Comparisons between at least two instruments are continuing in some of the current trials. A direct comparison of the QLQ-C30 with the MOS SF-36 is included in CO.7 and HN.1, while a comparison with the FACT-P (in PR.3) is being carried out by randomizing participating centers to using either the QLQ-C30 or the FACT-P.

Does HQL data provide supplementary data to standard toxicity data? It is standard practice to collect toxicity data in NCIC CTG clinical trials. The data are usually collected by clinical trials personnel (nurses or data managers) and reported in a standard format. However, does this information provide an accurate description of the impact that a given toxicity has on the patient's life? By collecting both toxicity and HQL data, it should be possible to compare the two methods. Preliminary data in one study (ME.7) suggest that more information is obtained from the HQL assessment than from the standard toxicity data (*12*). If this result is confirmed in further studies (e.g., LY.5 and SC.11), it will suggest that more attention should be paid to HQL data in the reporting of toxicity in the future.

Can study-specific modules be developed rapidly for phase III studies? Core HQL questionnaires, such as the QLQ-

Table 3. Completion rates for HQL questionnaires in NCIC CTG trials*

Trial symbol	Completion No. (time)	Expected	Received (%)	Complete/ received (%)
MA.5	1 (base line)	710	706 (99.4)	(76.3)
	2 (after cycle 1)	710	674 (94.9)	(93.5)
	3 (after cycle 2)	707	682 (96.5)	(94.3)
	4 (after cycle 3)	706	658 (93.2)	(97.1)
	5 (after cycle 4)	699	659 (94.3)	(97.4)
	6 (after cycle 5)	694	659 (95.0)	(97.0)
	7 (after cycle 6)	688	451 (65.6)	(95.1)
	8 (9 mo)	681	570 (83.7)	(94.9)
	9 (12 mo)	657	559 (85.1)	(94.1)
	10 (15 mo)	631	513 (81.3)	(94.3)
	11 (18 mo)	604	494 (82.2)	(94.7)
	12 (21 mo)	543	438 (80.7)	(92.0)
	13 (24 mo)	426	348 (81.7)	(92.8)
SC.8	1 (base line)	535	532 (99)	474 (89)
	2 (day 8)	535	491 (92)	431 (88)
	3 (day 15-28)	535	447 (84)	404 (90)
SC.9	1 (base line)	295	294 (99.7)	259 (88)
	2 (day 8)	295	298 (94)	237 (85)
	3 (day 15-28)	295	274 (93)	241 (88)
SC.11	1 (base line)	696	691 (99)	598 (86)
	2 (postchemotherapy)	696	655 (94)	594 (84)

*Complete compliance data for ME.7 has been published previously (*10*).

C30 and the FACT-G, are designed to be used in any population of patients with cancer. It has been recommended that modules of questions specific to disease sites or to individual studies be added to the core questionnaires (*13*). A method for the development of modules has been suggested by the EORTC Study Group on Quality of Life (*14*). An alternative to modules that contain domains is the checklist approach, in which each issue is treated as a single item (*15*).

The large number of clinical trials undertaken by the NCIC CTG has necessitated the rapid development of supplementary items to be used with the core questionnaires. Study-specific checklists have been added to the QLQ-C30, as listed in Table 4. Care has been taken to keep the checklists brief, so that the entire questionnaire package usually contains less than 50 items. In keeping with the QLQ-C30 response format, the checklist items are answerable in a four-category response option.

What is the significance of results from HQL assessments? The significance of results is usually expressed in statistical terms. However, with very large sample sizes, small numerical differences are often statistically significant at the $P<.05$ level. What is the impact of such small differences clinically; e.g., would they result in a clinical decision to alter the management of the patient's condition based on the result? This difference has been alluded to as the "minimal clinically important difference" (*16,17*). A variation of this concept is to ask what degree of change is perceived as being meaningful from the patient's perspective, i.e., "subjectively significant" or "subjectively meaningful" (*18*). To explore this approach, a subjective significance questionnaire was developed and is being used in several trials. Results are not yet available.

Discussion

The measurement of HQL in oncology has progressed rapidly in the last decade. Not only have new instruments been designed for use in populations of people with cancer, but many researchers and clinical trials groups in Australia, Europe, and North America are now incorporating HQL assessment in clinical trials. Furthermore, early difficulties with compliance reported in some clinical trials (*19,20*) appear to be lessening. These activities have yielded important lessons about the measurement of HQL in oncology (*21*).

Table 4. Study-specific modules and checklists

Brief title	Trial
Effect of chemotherapy-induced vomiting	SC.8, 9, 11, 12
Chemotherapy for metastatic breast cancer	MA.8, 10
Chemotherapy for colorectal cancer	CO.7, CO.9
Adjuvant treatment of malignant melanoma	ME.7
Radiation/chemotherapy for cervical cancer	CX.2, CX.3
Hormonal/radiation therapy for regional prostate cancer	PR.3, PR.5
Surgery/chemotherapy for non–small-cell lung cancer	BR.9
Chemotherapy/radiation therapy for non–small-cell lung cancer	BR.10
Chemotherapy for extensive-stage small-cell lung cancer	BR.8
Surgery for head and neck cancer	HN.1
Chemotherapy/radiation therapy for early-stage Hodgkin's disease	HD.6
Chemotherapy for advanced ovarian cancer	OV.10
Chemotherapy for non-Hodgkin's lymphoma in the elderly	LY.5
Chemotherapy for multiple myeloma	MY.7
Subjective significance of change in HQL	BR.8, 9, 10; CO.7, 9, 10; CX.3; MA.8; OV.10; PR.3; HD.6

The NCIC CTG has integrated HQL assessment in all but two of the clinical trials that it has initiated since 1989. This has been accepted by clinical investigators, nurses, and data managers to the point where it is now considered to be a routine part of a trial, analogous to the collection of laboratory, response, and survival data. The NCIC CTG also participates in intergroup trials initiated by the EORTC and North American clinical trials groups. However, in clinical trials initiated by other clinical trials groups that do not include HQL assessment, the CTG also does not assess HQL. It is anticipated that as HQL becomes a component in more trials initiated by other groups, the CTG will also measure HQL in those trials in which it participates. This will provide an opportunity to gain an even broader experience in more tumor sites and with more HQL instruments.

The NCIC CTG uses the EORTC QLQ-C30 (or variants thereof) in most of its clinical trials (21 of 27). The decision was made at the inception of HQL assessment that extensive experience with one instrument would lead to a thorough knowledge of its reliability and validity in a variety of circumstances and would allow cross-study comparisons and an opportunity to further the development of the instrument and ask additional research questions about the measurement of HQL. In retrospect, this decision seems to have been a reasonable one, and it is expected that recent data will lead to improvements in the reliability of the QLQ-C30 and the timing of HQL assessments in particular circumstances, and a better understanding of the significance of the results of HQL assessment.

In summary, the success of the NCIC CTG in implementing HQL assessment as an integral part of phase III clinical trials can be attributed to a commitment by the central office and clinical trials personnel to HQL measurement, the development of a policy for HQL assessment, the availability of writing guidelines for incorporation of HQL into protocols, and appropriate education of all personnel at a very early stage of implementation. In addition, it has been helpful to have concurrent meetings of the various disease site committees and data managers to provide frequent updates of progress and an opportunity for constructive suggestions from within the Quality-of-Life Committee.

References

(*1*) Osoba D. The Quality of Life Committee of the Clinical Trials Group of the National Cancer Institute of Canada: organization and functions. Qual Life Res 1992;1:211-8.

(*2*) Osoba D, Zee B, Sadura A, Pater J, Quirt I. Measurement of quality of life in an adjuvant trial of gamma interferon versus levamisole in malignant melanoma. In: Salmon SE, editor. Adjuvant therapy of cancer VII. Philadelphia: Lippincott, 1993:412-6.

(*3*) Levine MN, Guyatt GH, Gent M, De Pauw S, Hryniuk WM, Arnold A, et al. Quality of life in stage II breast cancer: an instrument for clinical trials [*see* comment citation in Medline]. J Clin Oncol 1988;6:1798-810.

(*4*) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365-76.

(5) Osoba D, Zee B, Pater J, Warr D, Kaizer L, Latreille J. Psychometric properties and responsiveness of the EORTC Quality of Life Questionnaire (QLQ-C30) in patients with breast, ovarian and lung cancer. Qual Life Res 1994;3:353-64.

(6) Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570-9.

(7) Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, et al. Measuring the quality of life in cancer patients: a concise QL-index for use by physicians. J Chronic Dis 1981;34:585-97.

(8) Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). 1. Conceptual framework and item selection. Med Care 1992;30:473-81.

(9) Osoba D, Warr D, Zee B, Kaizer L, Latreille J, Pater J. Pretreatment health-related quality of life (HQL) status predicts chemotherapy-induced emesis (CIE). Proc ASCO 1995;14:497.

(10) Sadura A, Pater J, Osoba D, Levine M, Palmer M, Bennett K. Quality-of-life assessment: patient compliance with questionnaire completion [*see* comment citation in Medline]. J Natl Cancer Inst 1992;8:1023-6.

(11) Niezgoda HE, Pater JL. A validation study of the domains of the core EORTC quality of life questionnaire. Qual Life Res 1993;2:319-25.

(12) Paul N, Pater JL, Whitehead M, et al. Methods of toxicity collection: an evaluation of the relative effectiveness of the case report flow sheet, the patient symptom diary, and the quality of life questionnaire. Control Clin Trials 1991;12:648.

(13) Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality-of-life assessment in cancer clinical trials. Recent Results Cancer Res 1988;111:231-49.

(14) Sprangers MA, Cull A, Bjordal K, Groenwald M, Aaronson NK. The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. Qual Life Res 1993;2:287-95.

(15) Osoba D. Self-rating symptom checklists: a simple method for recording and evaluating symptom control in oncology. Cancer Treat Rev 1993;10 Suppl A:43-51.

(16) Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.

(17) Juniper EF, Guyatt GH, Willian A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81-7.

(18) Osoba D. Measuring the effect of cancer on health-related quality of life. PharmacoEconomics 1995;7:308-19.

(19) Ganz PA, Haskell CM, Figlin RA, La Soto N, Siau J. Estimating the quality of life in a clinical trial of patients with metastatic lung cancer using the Karnofsky performance status and the Functional Living Index—Cancer. Cancer 1988;61:849-56.

(20) Hürny C, Bernhard J, Joss R, Willems Y, Cavalli F, Kiser J, et al. Feasibility of quality of life assessment in a randomized phase III trial of small cell lung cancer—a lesson from the real world. Ann Oncol 1992;3:825-31.

(21) Osoba D. Lessons learned from measuring health-related quality of life in oncology [*see* comment citation in Medline]. J Clin Oncol 1994;12:608-16.

PARTICIPANT LIST

Jeffrey Abrams
CIB/CTEP/NCI
6130 Executive Blvd.
EPN 741
Bethesda, MD 20892
(301) 496-4844
Fax (301) 402-0557

Barrie Anderson
University of Iowa/GOG
4630 JCP, University of Iowa Hospitals
and Clinics
Iowa City, IA 52242
(319) 356-2015
Fax (319) 353-8363

Linda Anderson
National Cancer Institute
Bldg. 31, Rm. 10A19
Bethesda, MD 20892
(301) 496-6641
Fax (301) 496-0846

Susan G. Arbuck
IDB/CTEP/NCI
Executive Plaza North, Rm. 715
Bethesda, MD 20892
(301) 496-1196
Fax (301) 402-0428

Frank Baker
Johns Hopkins School of Hygiene
and Public Health
615 N. Wolfe Street, Ste 7513
Baltimore, MD 21205
(410) 955-4074
Fax (410) 955-1811

Julie Baltz
NCI/DCT/CTEP/PMB
EPN Rm. 804
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-5725
Fax (301) 402-4870

Ivan Barofsky
Johns Hopkins Bayview Medical Center
4940 Eastern Ave.
Baltimore, MD 21224
(410) 550-0162
Fax (410) 558-9346

Julie Beitz
FDA/Center for Drug Evaluation and Research
5600 Fishers Lane, HFD-150
Rockville, MD 20857
(301) 594-5745
Fax (301) 594-0498

Kathy Benjamin
AHCPR
2101 E. Jefferson St.
Rockville, MD 20852
(301) 594-1485
Fax (301) 594-3211

Amy Bonomi
Rush Cancer Institute
Rush-Presbyterian-St. Luke's Medical Center
Ste. 863
1725 W. Harrison
Chicago, IL 60612
(312) 850-8934
Fax (312) 850-8931

Drew Bradlyn
Pediatric Oncology Group &
West Virginia University
Dept. of Behavioral Med.
930 Chestnut Ridge Rd.
Morgantown, WV 26505
(304) 293-2411
Fax (304) 293-5555
(304) 293-8724

Mark F. Brady
GOG Statistical Office
Elm & Carlton St.
Buffalo, NY 14221
(716) 845-5702
Fax (716) 845-8393

Nancy Breen
NCI
EPN Rm. 313
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-4675
Fax (301) 402-0816

Christine Carter
The EMMES Corporation
11325 Seven Locks Rd.
Potomac, MD 20854
(301) 299-8655
Fax (301) 299-3991

Shelly Carter-Campbell
The EMMES Corporation
11325 Seven Locks Rd.
Potomac, MD 20854
(301) 299-8655
Fax (301) 299-3991

David Cella
Rush Cancer Institute/ECOG
Rush-Presbyterian-St. Luke's Medical Center
Ste. 863
1725 W. Harrison
Chicago, IL 60612
(312) 563-2410
Fax (312) 563-2471

T. Timothy Chen
BRB/CTEP/DCT/NCI
Executive Plaza North, Rm. 739
Bethesda, MD 20892
(301) 402-0640
Fax (301) 402-0560

Isagani Mario S. Chico
CTEP/DCT/NCI
6130 Executive Blvd., EPN 715
Bethesda, MD 20892
(301) 496-1196
Fax (301) 402-0428

Rena Convissor
Center for the Advancement of Health
2000 Florida Ave., NW #210
Washington, DC 20009
(202) 387-2829
Fax (202) 387-2854

Miles Cooper
The Comprehensive Cancer Center
of Wake Forest University
Medical Center Blvd.
Winston-Salem, NC 27157-1082
(910) 716-4300
Fax (910) 716-5687

Katherine Crosson
NCI Bldg. 31, Rm. 10A10
9000 Rockville Pike
Bethesda, MD 20895
(301) 496-6792
Fax (301) 496-7063

Barbara Curbow
Johns Hopkins University
624 N. Broadway, Rm. 705
Baltimore, MD 21205
(410) 955-2312
Fax (410) 955-7241

Janet Dancey
NCIC Clinical Trials Group
Queen's University
18 Barrie St.
Kingston, Ontario, Canada K7L3N6
(613) 545-6430
Fax (613) 545-2411

Richard Day
University of Pittsburgh
Graduate School of Public Health
Department of Biostatistics
Pittsburgh, PA 15761
(412) 624-3032
Fax (412) 624-9969

Haim Erder
AMGEN, Inc.
1840 DeHavilland Dr.
Tower Oaks, CA 91320
(805) 447-6670
Fax (805) 498-0358

Pennifer Erickson
National Center for Health Statistics
6525 Belcrest Rd. #730
Hyattsville, MD 20782
(301) 436-5975
Fax (301) 436-3572

Carol Estwing Ferrans
University of Illinois at Chicago
College of Nursing (M/C 802)
45 S. Damen Ave.
Chicago, IL 60612
312) 996-7900
Fax (312) 996-4979

Diane L. Fairclough
ECOG Statistical Center
4 Binney St., Meyer 428
Boston MA 02115
617) 632-2439
Fax (617) 632-2444

Ellen Feigal
CIB/CTEP/NCI
130 Executive Blvd., EPN 741
Bethesda, MD 20892
301) 496-4844
Fax (301) 402-0557

John Ferguson
OD/ODP/OMAR/NIH
Federal Bldg., 618
550 Wisconsin Ave.
Bethesda, MD 20814
301) 496-5641
Fax (301) 402-0420

Lou Fintor
NCI
130 Executive Blvd., EPN 313
Rockville, MD 20852
301) 496-8500
Fax (301) 496-8667

John F. Foley
University of Nebraska Med. Ctr.
00 South 42nd St.
Omaha, NE 68198-3330
402) 559-6313
Fax (402) 559-6520

Steven Fox
Agency for Health Care Policy and Research
101 East Jefferson St., Rm. 605
Rockville, MD 20852
301) 594-1485
Fax (301) 594-3211

Patricia A. Ganz
UCLA Schools of Medicine & Public Health
Division of Cancer Prevention &
Control Research
100 Glendon Ave., Ste. 711
Los Angeles, CA 90049
310) 206-1404
Fax (310) 206-3566

Kathie Garrett
Cancer Survivorship Programs
AMC Cancer Research Ctr.
1600 Pierce St.
Denver, CO 80013
(303) 239-3420
Fax (303) 233-1863

Clare Gnecco
Food and Drug Administration
5600 Fisher's Lane
Rockville, MD 20857
(301) 594-5774
Fax (301) 594-0498

Betty L. Goon
R.W. Johnson Pharmaceutical Research Institute
P.O. Box 300
700 Route 202s
Raritan, NJ 08869
(908) 704-4576
Fax (908) 526-1242

Carolyn C. Gotay
Cancer Research Center of Hawaii
1263 Lauhala St., Rm. 406
Honolulu, HI 96813
(808) 586-2975
Fax (808) 586-3016

Nancy L. Hedlund
Cancer Research Center of Hawaii
1236 Lauhala St., Rm. 406
Honolulu, HI 96813
(808) 586-2975
Fax (808) 586-3016

James E. Herndon II
CALGB Statistical Center
2024 W. Main St., Ste. B-101
Durham, NC 27705
(919) 416-5108
Fax (919) 286-3956

Penelope Hopwood
CRC Psycological Medicine Group
Christie Hospital
Stanley House
Wilmslow Rd.
Withington, Manchester, U.K.
M204BX
0161-446-3679
Fax 0161-448-1655

Karen Iseminger
GOG Statistical Office
Roswell Park Cancer Institute
Elm & Carlton Sts.
Buffalo, NY 14263-0001
(716) 845-8876
Fax (716) 845-8393

Lee Ann Jensen
National Heart, Lung, and Blood Institute
5104 Federal Bldg.
7550 Wisconsin Ave.
Bethesda, MD 20892
(301) 496-8387
Fax (301) 402-1622

David Jodrey
Johns Hopkins School of Hygiene and Public Health
615 N. Wolfe St.
Baltimore, MD 21205
(410) 955-4074
Fax (410) 955-1811

Richard Kaplan
CIB/CTEP/NCI
6130 Executive Blvd., EPN 741
Bethesda, MD 20892
(301) 496-2522
Fax (301) 402-0557

Gwendoline M. Kiebert
EORTC
Ave. E Mounier 83, Box 11
1200 Brussels, Belgium
32-2-774-1611
Fax 32-2-772-3545

Jacek Kopec
NSABP
230 McKee Place, Ste. 402
Pittsburgh, PA 15213
(412) 383-1420
Fax (412) 383-1388

Edward Korn
BRB/CTEP/NCI
6130 Executive Blvd., EPN 739
Bethesda, MD 20892
(301) 496-4836
Fax (301) 402-1560

Alice B. Kornblith
Memorial Sloan-Kettering Cancer Ctr.
CALGB
1275 York Ave.
New York, NY 10021
(212) 639-3278
Fax (212) 752-7185

Laura Loll
FHCRC-SWOG
1124 Columbia St., MP-557
Seattle, WA 98104-2092
(206) 667-4846
Fax (206) 667-4408

Sally Lopez
The EMMES Corporation
11325 Seven Locks Rd.
Potomac, MD 20854
(301) 299-8655
Fax (301) 299-3991

June Lunney
National Institute of Nursing Research
Bldg. 45, Rm. 3AN-12
Bethesda, MD 20892-6300
(301)594-6908
Fax (301) 480-8260

David Maehin
MRC Cancer Trials Office
5 Shaftesbury Rd.
Cambridge, UK
CB2 2BW
011-44-1223-311110
(Fax) 011-44-1223-311844

William E. MacLean Jr.
Peabody College of Vanderbilt University/CCG
Box 521
Nashville, TN 37203
(615) 322-8141
Fax (615) 343-9494

Alfred C. Marcus
AMC Cancer Research Ctr.
1600 Pierce St.
Denver, CO 80214
(303) 239-3397
Fax (303) 233-1863

Mary McCabe
IDB/CTEP/DCT/NCI
6130 Executive Blvd., EPN Rm.. 715
Bethesda, MD 20892
(301) 496-5223
Fax (301) 402-0428

Richard P. McQuellon
Comp. Cancer Ctr of Wake Forest U.
Bowman Gray School of Medicine
Medical Ctr. Blvd.
Winston-Salem, NC 27157-1082
(910) 716-4102
Fax (910) 716-5687

Beth Meyerowitz
University of Southern California
Dept. of Psychology
Los Angeles, CA 90089-1061
(213) 740-2209
Fax (213) 746-0048

Carol M. Moinpour
Southwest Oncology Group Statistical Ctr
FHCRC MP 557
1124 Columbia St.
Seattle, WA 98104
(206) 667-4604
Fax (206) 667-4408

Elizabeth Moore
RAB/CTEP/DCT/NCI
EPN Rm. 718
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-7912
Fax (301) 402-1584

Gray R. Morrow
URCC CCop Research Base
601 Elmwood Ave., Box 704
Rochester, NY 14642
(716) 275-5513
Fax(716) 273-1042

Claudia S. Moy
Johns Hopkins University
550 N. Broadway
Baltimore, MD 21205
(410) 955-8943
Fax (410) 955-0569

Anita Nelson
The EMMES Corporation
11325 Seven Locks Rd.
Potomac, MD 20892
(301) 299-8655
Fax (301) 299-3991

Cherie Nichols
NCI
Bldg. 31, Room 11A21
Bethesda, MD 20892
(301) 496-5515
Fax (301) 402-1225

Robert B. Noll
Children's Hospital Medical Ctr./CCG
University of Cincinnati
Division of Hematology/Oncology
3333 Burnet Ave.
Cincinnati, OH 45229-3039
(513) 559-4266
Fax (513) 559-3549

David Osoba
Communities Oncology Program
BC Cancer Agency
600 West 10th Ave.
Vancouver, BC, Canada
V52 4E6
(604) 877-6000
Fax (604) 872-4596

Rose Mary Padberg
NCI
EPN Rm. 300
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-8541
Fax (301) 496-8667

Richard Payne
M. D. Anderson Cancer Center
1515 Holcombe Blvd.
Harrison, TX 73012
(713) 794-2824
Fax (713) 794-4999

Pam Phillips
The EMMES Corporation
11325 Seven Locks Rd.
Potomac, MD 20854
(301) 299-8655
Fax (301) 299-3991

Sandra L. Pineros
Fred Hutchinson Cancer Research Ctr.
SWOG
10530 59 Ave. W.
Mukilteo, WA 98275
(206) 349-1130
Fax (206) 667-4408

Brad H. Pollock
Pediatric Oncology Group Statistical Office
Dept. of Health Policy & Epidemiology
Univ. of Florida College of Medicine
104 N. Main St., Ste. 600
Gainesville, FL 32601-3330
(904) 392-5198
Fax (904) 392-8162

Sheila Prindiville
DCPT/NCI
EPN 300F
Bethesda, MD 20892
(301) 496-8541
Fax (301) 496-8667

Leslie L. Robison
University of Minnesota
420 Delaware St. SE, Box 422
UMHC
Minneapolis, MN 55455
(612) 626-2778
Fax (612) 626-2815

Esther Rose
R.W. Johnson Pharmaceutical
Research Institute
PO Box 300
700 Route 202S
Raritan, NJ 08869
(908) 704-4576
Fax (908) 526-1242

Julia H. Rowland
Georgetown University Med. Ctr.
3800 Reservoir Rd. NW
Washington, DC 20007
(202) 687-6528
Fax (202) 687-6658

Mira Rubenstein
AMC Cancer Research Ctr.
1600 Pierce St.
Denver, CO 80013
(303) 239-3425
Fax (303) 233-1863

Harland Sather
Childrens Cancer Group
440 E. Huntington Drive, Ste. 300
PO Box 60012
Arcadia, CA 91066-6012
(818) 447-0064
Fax (818) 445-4334

Patricia R. Schettino
PMB/CTEP/NCI
Executive Plaza North
Rm. 804
6130 Executive Blvd.
Rockville, MD 20852
(301) 496-5725
Fax (301) 402-4870

Charles Scott
Radiation Therapy Oncology Group
1101 Market St., 14th Floor
Philadelphia, PA 19107
(215) 574-3208
Fax (215) 928-0153

Malcolm Smith
CIB/CTEP/DCT/NCI
EPN Rm. 741
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-4884
Fax (301) 402-0557

Vera Suman
Mayo Clinic
Plummer 4, First St., SW
Rochester, MN 55905
(507) 284-8803
Fax (507) 284-1902

Karen Syrjala
Fred Hutchinson Cancer Research Center
1124 Columbia St., FB635
Seattle, WA 98104
(206) 667-4579
Fax (206) 667-3531

Edward L. Trimble
CIB/CTEP/DCT/NCI
6130 Executive Blvd., EPN 741
Bethesda, MD 20892
(301) 496-2552
Fax (301) 402-0557

Richard Ungerleider
CIB/CTEP/DCT/NCI
EPN Rm. 741
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-6056
Fax (301) 402-0557

Claudette Varricchio
Community Oncology and Rehabilitation Branch
6130 Executive Blvd., EPN 300
Bethesda, MD 20892
(301) 496-8541
Fax (301) 496-8667

Judith Wagner
U.S. Congress OTA
Washington, DC 20510
(202) 228-6590
Fax (202) 228-6603

Richard B. Warnecke
Survey Research Lab
University of Illinois at Chicago
910 W. Van Buren St., Ste 505
(M/C336)
Chicago, IL 60093
(312) 996-6130
Fax (312) 996-3358

Todd Wasserman
Radiation Oncology Ctr.
4939 Children's Pl., Ste 5500
St. Louis, MO 63110
(314) 362-8501
Fax (314) 362-8521

Deborah Watkins Bruner
Fox Chase Cancer Center
RTOG
422 Park Ave.
Swarthmore, PA 19081
(610) 544-7191
Fax (610) 544-9084

Jane Weeks
Dana-Farber Cancer Inst.
44 Binney St.
Boston, MA 02115
(617) 632-2509
Fax (617) 632-3161

Rodger J. Winn
M. D. Anderson Cancer Center
1515 Holcombe Blvd.
Box 501
Houston, TX 77030
(713) 792-8515
Fax (713) 796-9155

Rosemary Yancik
Liaison and Applied Research
National Institute on Aging, NIH
Bldg. 31, Rm. 5C05
Bethesda, MD 20892-2292
(301) 496-5278
Fax (301) 496-2793

Benny Zee
Clinical Trials Group, NCIC
82-84 Barrie St.
Kingston, Ontario, Canada
K7L 3N6
(613) 545-6430
Fax (613) 545-2941

NIH LIBRARY

JOURNAL OF THE NATIONAL CANCER
INSTITUTE MONOGRAPHS

NO.21 1996 MISSING

MONOGRAPHS
JOURNAL OF THE NATIONAL CANCER INSTITUTE

NATIONAL CANCER INSTITUTE

National Institutes of Health Consensus Conference on Breast Cancer Screening for Women Ages 40–49

1997
Number 22

Contents

(Contents continued on back cover)

MONOGRAPHS

JOURNAL OF THE NATIONAL CANCER INSTITUTE

Number 22
ISSN 0027-8874
ISBN 0-19-922326-2

1997

EDITORIAL POLICY: Manuscripts from key conferences dealing with cancer and closely related research fields, or a related group of papers on specific subjects of importance to cancer research, are considered for publication, with the understanding that they have not been published previously and are submitted exclusively to the *Journal of the National Cancer Institute Monographs.* All material submitted for consideration will be subject to review, when appropriate, by at least one outside reviewer and one member of the Editorial Board of the *Journal of the National Cancer Institute.* Opinions expressed by the authors are not necessarily those of the publisher or the editors.

Proposals for monographs should be submitted to the Editor-in-Chief, *Journal of the National Cancer Institute,* National Cancer Institute, 7550 Wisconsin Avenue, Room 114, Bethesda, MD 20814.

Journal of the National Cancer Institute Monographs are available on request to members of the National Cancer Institute Information Associates Program (one copy of each monograph per member). Monographs are also available from Oxford University Press, Journals Subscription Department, 2001 Evans Road, Cary, NC 27513 USA. Telephone toll-free in the U.S. and Canada, 1-800-852-7323, or 919-677-0977. Fax: 919-677-1714. E-mail: jnlorders@oup-usa.org.

National Institutes of Health Consensus Conference on Breast Cancer Screening for Women Ages 40–49

Proceedings of a Conference
Held at the
National Institutes of Health
Bethesda, Maryland
January 21–23, 1997

Conference Sponsors

National Cancer Institute
Richard Klausner, M.D.
Director

Office of Medical Applications of Research, NIH
John H. Ferguson, M.D.
Director

Conference Cosponsors

National Institute on Aging
Richard J. Hodes, M.D.
Director

Office of Research on Women's Health, NIH
Vivian W. Pinn, M.D.
Director

Centers for Disease Control and Prevention
David Satcher, M.D., Ph.D.
Director

National Institutes of Health Consensus Conference on Breast Cancer Screening for Women Ages 40–49

1997
Number 22

Contents

National Institutes of Health Consensus Development Conference Statement: Breast Cancer Screening for Women Ages 40-49, January 21-23, 1997

*National Institutes of Health Consensus Development Panel**

Foreword

The National Institutes of Health (NIH) Consensus Development Program, managed by the Office of Medical Applications of Research, is a unique technology assessment process in American medicine and is designed to produce a consensus statement at the end of a 3-day consensus conference. A consensus statement is a thoughtful and thorough data-driven synthesis of the current science based on a comprehensive review of the existing peer-reviewed medical literature, a series of state-of-the-art scientific presentations, and public testimony. The resulting statement helps to advance and clarify the field of science it addresses and provides an important and useful public health message.

The existence of controversy is a major criterion for determining the need to conduct an NIH consensus development conference. As such, there may be times when a panel cannot reach a consensus or when the panel's consensus is that there is no consensus. All NIH consensus panels are offered the opportunity to make a minority statement if a consensus cannot be obtained. In the previous 102 consensus conferences held by NIH over the past 20 years, this has happened on only two occasions.

This NIH Consensus Statement on Breast Cancer Screening for Women Ages 40-49 contains two reports: a majority report and a minority report. While a consensus was initially achieved by the entire panel at the end of the consensus conference, 2 of the 12 panel members subsequently differed on specific issues in the draft document in the weeks that followed and, ultimately, did not agree entirely with the majority statement.

The panel members writing the majority report took into consideration the risks versus the benefits of mammography and did not think that the data supported a recommendation for universal mammography screening for all women in their forties. The authors of the minority report believed the risks to be overemphasized by the majority and concluded the data did support a recommendation for mammography screening for all women in this age group. The entire panel did agree that women and their health care providers should be provided information on these issues upon which to base their decisions. Additionally, all panelists agreed that, for women in their forties who choose to have mammography, the costs of mammograms should be reimbursed by third-party payors or covered by health maintenance organizations.

It is in the spirit of providing all views on this controversial topic that both the majority and minority statements are presented.

John H. Ferguson, M.D., Director
Office of Medical Applications of Research
National Institutes of Health

Abstract

***Objective:* To provide health care providers, patients, and the general public with a responsible assessment of currently available data regarding the effectiveness of mammography screening for women ages 40-49. *Participants:* A non-Federal, nonadvocate, 12-member panel representing the fields of oncology, radiology, obstetrics and gynecology, geriatrics, public health, and epidemiology and including patient representatives. In addition, 32 experts in oncology, surgical oncology, radiology, public health, and epidemiology, presented data to the panel and to a conference audience of 1,100. *Evidence:* The literature was searched through Medline and an extensive bibliography of references was provided to the panel and the conference audience. Experts prepared abstracts with relevant citations from the literature. Scientific evidence was given precedence over clinical anecdotal experience. *Consensus Process:* The panel, answering predefined questions, developed its conclusions based on the scientific evidence presented in open forum and the scientific literature. The panel composed a draft statement that was read in its entirety and circulated to the experts and the audience for comment. Thereafter, the panel resolved conflicting recommendations and released a revised draft statement at the end of the conference. The final statement with a minority report was completed within several weeks after the conference. *Conclusions:* The Panel concludes that the data currently available do not warrant a universal recommendation for mammography for all women in their forties. Each woman should decide for herself whether to undergo mammography. Her decision may be based not only on an objective analysis of the scientific evidence and consideration of her individual medical history, but also on how she perceives and weighs each potential risk and benefit, the values she places on each, and how she deals with uncertainty. However, it is not sufficient just to advise a woman to make her own decision about mammograms. Given both the importance and the complexity of the issues involved in assessing the evidence, a woman should have access to the best possible relevant information regarding both benefits and risks, presented in an understandable and usable form. Information should be developed for women in their forties regarding potential benefits and risks to be provided to enable each woman to make the most appropriate decision. In addition, educational material to accompany this information should be prepared that will lead women step by step through the process of using such information in the best possible way for reaching a decision. For women in their**

*The members of the Consensus Development Panel are the authors of this manuscript (*see* page xiii).

Correspondence to: Bill Hall, Office of Medical Applications of Research, National Institutes of Health, Federal Bldg., Rm. 618, Bethesda, MD 20892. E-mail: billhall@nih.gov

Reproduced from J Natl Cancer Inst 1997;89:1015–26.

forties who choose to have mammography performed, the costs of the mammograms should be reimbursed by third-party payors or covered by health maintenance organizations so that financial impediments will not influence a woman's decision. Additionally, a woman's health care provider must be equipped with sufficient information to facilitate her decisionmaking process. Therefore, educational material for physicians should be developed to assist them in providing the guidance and support needed by the women in their care who are making difficult decisions regarding mammography. The two panel members writing a minority report believed the risks of mammography to be overemphasized by the majority and concluded that the data did support a recommendation for mammography screening for all women in this age group and that the survival benefit and diagnosis at an earlier stage outweigh the potential risks. [Monogr Natl Cancer Inst 1997;22:vii-xviii]

Introduction

Breast cancer is the single leading cause of death for women ages 40-49 in the United States. A 40-year-old woman has a 2 percent chance of being diagnosed with invasive breast cancer or ductal carcinoma *in situ* in the next 10 years, and her chance of dying from breast cancer during this decade is 0.3 percent. In addition to morbidity and mortality from breast cancer itself, women must endure the emotional impact of both the disease and its treatment, as well as the fear engendered by the threat of the disease.

To what extent can early detection through mammographic screening reduce the impact of breast cancer in women in their forties, and what risks may be associated with mammography in this age group? Although nonrandomized observational data on women screened with mammography have been reported, the benefits and risks of mammography screening for women in their forties can be validly assessed only by analyzing results obtained from clinical trials in which women are randomly assigned to be screened or not screened. A number of randomized clinical trials in 50- to 69-year-old women have shown clearly that early detection of breast cancer by mammography at regular intervals, with and without clinical breast examination (CBE), reduces breast cancer mortality by about one-third. However, the results have not been as clear for women ages 40-49. Internationally, experts have continued to examine data regarding the use of mammography in this age group. Results of several trials in different countries have been updated recently with longer periods of observation.

To address this issue and to examine newly available data from both observational studies and randomized trials, the National Cancer Institute, together with the Office of Medical Applications of Research of the National Institutes of Health (NIH), convened a Consensus Development Conference on Breast Cancer Screening for Women Ages 40-49. The conference was cosponsored by the National Institute on Aging, the NIH Office of Research on Women's Health, and the Centers for Disease Control and Prevention. Following a day and a half of presentations by experts in the relevant fields and discussion from the audience, an independent consensus panel composed of specialists and generalists (including epidemiologists, statisticians, radiologists, and oncologists), representatives from the public, and other experts considered the evidence and formulated a consensus statement in response to the following five predefined questions:

- Is there a reduction in mortality from breast cancer due to screening women ages 40-49 with mammography, with or without physical examination? How large is the benefit? How does this change with age?
- What are the risks associated with screening women ages 40-49 with mammography and with or without physical examination? How large are the risks? How do they change with age?
- Are there other benefits? If so, what are they? How do they change with age?
- What is known about how the benefits and risks of breast cancer screening differ based on known risk factors for breast cancer?
- What are the directions for future research?

1) Is There a Reduction in Mortality From Breast Cancer Due to Screening Women Ages 40-49 With Mammography, With or Without Physical Examination? How Large Is the Benefit? How Does This Change With Age?

Information regarding the usefulness of screening procedures is provided by randomized controlled trials (RCTs) in which participants are randomly assigned to receive or not receive screening. Currently available data from eight RCTs that included women ages 40-49 have been used to examine the effect of screening mammography on breast cancer mortality. Such studies must include long-term follow-up in order to account for the variable course of breast cancer and to examine the ultimate benefit—a reduction in mortality from breast cancer. In fact, the benefit of reduced breast cancer mortality in the summary of these studies is about half that seen in women ages 50-69. About twice as much follow-up time is needed to see the benefits.

These trials were begun between 1963 and 1982. On the basis of a summary of data from these RCTs, there is no statistically significant difference in breast cancer mortality within 7 years after screening is initiated between women randomized to receive or not receive screening. Summary data in five of eight RCTs show a trend toward reduced breast cancer mortality only after a follow-up of 10 or more years, with the decrease estimated at 16 percent (with confidence intervals from 2 percent to 28 percent). In the RCTs, many of the women began mammography while they were in their late forties and continued to have mammography after age 50. Consequently, one cannot determine if the women who benefited from mammography in these studies showed this benefit because of breast cancer diagnosis following mammographic screening performed after age 50.

Based on meta-analyses of the RCTs, regular screening of 10 000 women ages 40-49 would result in extension of the lives of 0-10 women. About 2,500 women would have to be screened regularly in order to extend one life. For those women whose survival is extended, the length of life extension is not known.

The magnitude of the benefit seen in the RCTs may be underestimated for several reasons. First, only one of these trials was specifically designed to study women in their forties. Second, in all the trials, some women assigned to screening were not screened, and some assigned to the control group obtained screening outside the trial. Third, trials varied in the length of the screening interval used, ranging from 1 to 2 years, which may be

too long to detect fast-growing cancers before they become clinically evident. Finally, current mammographic technology has improved in the past 15 years from that used in the RCTs initiated between 1963 and 1982. Many of these same factors operate in RCTs of women ages 50-69 years, so that the benefits could also have been underestimated in older women.

The incidence of breast cancer approximately doubles from ages 40-44 to ages 45-49. This increased incidence suggests that any benefit of mammography in women ages 40-49 may be greater for women in their late forties. Because a disproportionate number of women in the screening phase of these trials were in their late forties, it is difficult to assess the relative benefits of mammography for the younger women within the 40- to 49-year-old group compared with the older women.

In addition to RCTs, uncontrolled case series comparing women with mammographically detected breast cancer to women with clinically detected cancers show that mammography finds breast cancers at an earlier stage. Earlier stage cancers generally have better prognoses. However, it is not necessarily valid to conclude that screening mammography results in fewer breast cancer deaths, because screening selectively identifies women with slow-growing cancers whose prognosis is better, regardless of treatment. Detection at an earlier stage is relevant only if it can be shown in a randomized study that fewer deaths occur in a screened population than in a comparable unscreened control population.

2) What Are the Risks Associated With Screening Women Ages 40-49 With Mammography and With or Without Physical Examination? How Large Are the Risks? How Do They Change With Age?

Understanding the nature and magnitude of risks is important to both primary care providers and women making informed decisions about breast cancer screening. Critical issues include the following: risks associated with false-negative examinations, additional diagnostic testing induced by false-positive examinations, psychosocial consequences of abnormal examinations, potential risk of overtreatment of low-risk or *in situ* cancers, and potential risk from radiation exposure.

False-negative mammograms. Up to one-fourth of all invasive breast cancers are not detected by mammography in 40- to 49-year-olds, compared with one-tenth of cancers in 50- to 69-year-olds. Women with these cancers may be harmed if their diagnosis or treatment is delayed because of a normal, or false-negative, mammogram. Professional and public education as well as disclaimers on mammography reports have increased the awareness of this problem in women with clinical symptoms, but more attention should be given to the issue in screened women.

False-positive mammograms. Many mammographic abnormalities may not be cancer but will prompt additional testing and anxiety. Approximately 10 percent of all screening mammograms are read as abnormal, each of which will prompt the performance of an average of two additional diagnostic tests such as diagnostic mammography, ultrasound, needle aspiration, core biopsy, or surgical biopsy. Given the lower incidence of breast cancer in 40- to 49-year-old women compared with that in older women, false-positive examinations are more common in younger women, and the proportion of true-positive examinations increases with increasing age. As many as 3 out of 10 women who begin annual screening at age 40 will have an abnormal mammogram during the next decade. For women ages 40-49 undergoing breast biopsy for mammographic findings, only half as many cancers are diagnosed compared with women ages 50-69. For every eight biopsies performed in the younger age group, one invasive and one *in situ* breast cancer are found.

Psychosocial consequences. There is concern that women who have abnormal mammograms—both true-positive and false-positive—experience psychosocial sequelae, including anxiety, fear, and inconvenience. Additional information is needed on whether experiencing a false-positive mammogram may affect subsequent willingness to undergo future screening mammography at ages when it is of greatest benefit.

Low-risk cancer and ductal carcinoma *in situ*. Not all women diagnosed with breast cancer by mammographic screening are helped by early detection. Some have slow-growing cancers that may be successfully treated when discovered later. Some cancers that might be detected in women in their forties are so slow growing that they could be detected by mammograms after age 50 and treated at that time. Earlier detection may cause additional months or years of cancer-related anxiety, affecting personal and workplace relationships, as well as insurance coverage.

Ductal carcinoma *in situ* (DCIS) is frequently diagnosed in mammographically screened women ages 40-49. DCIS is a heterogeneous entity for which the natural history, clinical significance, prognostic factors, and treatment are uncertain. Because some cases of DCIS may not progress to invasive cancer, a risk of overtreatment exists.

Radiation exposure. The risk of radiation-induced breast cancer has long been a concern to mammographers and has driven the efforts to reduce the radiation dose per examination. Radiation has been shown to cause breast cancer in women, and the risk is proportional to dose. The younger the woman at the time of exposure, the greater her lifetime risk for breast cancer. Radiation-related breast cancers occur at least 10 years after exposure. However, breast cancer as a result of the radiation dose associated with mammography has not been demonstrated. Radiation from yearly mammograms during ages 40-49 has been estimated as possibly causing 1 additional breast cancer death per 10 000 women. However, this estimate is based on statistical models from epidemiological studies of high-dose exposures, and the actual risk at the lower doses associated with mammography could range from much higher than one to nonexistent. Women with inherited or acquired defects in DNA repair mechanisms may have a different susceptibility to the effects of radiation.

3) Are There Other Benefits? If So, What Are They? How Do They Change With Age?

Additional benefits from screening women ages 40-49 may include earlier detection and increased compliance. Data from several studies suggest that the average size of newly diagnosed breast cancer is decreasing and that the proportion of stages 0 and I cancers (i.e., DCIS and small invasive breast cancer) is increasing due to mammographic screening in women ages 40-49. The increased detection of DCIS may prove beneficial if it leads to a subsequent decrease in the incidence of invasive cancer. This increased detection and treatment of early-stage cancer

or premalignant changes could be consistent with a reduction in breast cancer mortality appearing only after 10 years following the initiation of screening.

The diagnosis of breast cancer at a smaller size or earlier stage will allow a woman more choice in selecting among various treatment options. For example, more women with cancer detected by mammography have the option of lumpectomy, rather than mastectomy, compared with women whose cancers are detected by palpation. Studies also show that the rate of axillary dissection or chemotherapy may be reduced among women who have smaller or earlier stage cancer. This choice in type of treatment allows a woman a measure of control over treatment decisions. The value of this benefit must be individually assessed.

Bringing women into screening programs at a younger age could provide an earlier opportunity for patient education and increase their access to and utilization of health care. However, there is no information on whether initiating mammographic screening at age 40 would increase or decrease screening compliance in later years.

Women with true-negative mammogram screening tests may benefit from reassurance that they do not have breast cancer. However, the reassurance value of a true-negative screen has not been studied and is complicated by the fact that it is not possible to distinguish true negatives from false negatives without additional testing.

4) What Is Known About How the Benefits and Risks of Breast Cancer Screening Differ Based on Known Risk Factors for Breast Cancer?

Although much is known about risk factors for breast cancer incidence and mortality, little is known about the effects of screening in high-risk subgroups. Known risk factors include family history of breast cancer, having no children, and having a first birth after age 30. None of the RCTs of breast cancer screening for women in their forties has examined the effect of screening on the mortality of women in any of the high-risk subgroups. Most of these trials included only white women. Although the incidence of breast cancer is the same for African-American women and white women in their forties, African-American women have a 50 percent higher breast cancer mortality rate than white women in this age group. An outreach screening program enrolling a large number of women from minority groups has reported that Hispanic and Native-American women have higher false-positive rates than white women in their forties. A practice-based screening program including women ages 40-49 found a higher cancer detection rate and a lower false-positive rate for women with a family history of breast cancer.

5) What Are the Directions for Future Research?

There are insufficient data to address several aspects of screening mammography. Although the focus of this conference has been specifically on women ages 40-49, future research should examine the effects of mammography for all ages at risk. Age is a continuum; although one can use an artificial cutoff of 50 as an approximation of the age of menopause and its associated biologic changes, age should be studied as a continuum. The ongoing UK-AGE and Eurotrials may add valuable information on benefits and risks of screening specifically in this age group.

Most of the following research questions should be answered for women of all ages:

1. What is the optimum screening interval for women of various ages?
2. How much of the mortality benefit found in the RCTs among women ages 40-49 can be explained by factors other than mammographic screening—for example, by screening at later age or by improved treatment?
3. How does the mortality reduction for women depend on the age at which screening mammography begins?
4. Will women receive more or less radiation therapy or chemotherapy because of early detection of breast cancer? What are the consequences of these treatments?
5. What are the psychosocial benefits and risks of mammography?
6. Would initiating mammographic screening at age 40 increase screening compliance in later years? Would it provide an opportunity for education regarding prevention services and use of health care?
7. Does the benefit or risk of mammography differ by race or ethnicity? If the benefit is less, are there adjunctive measures that could improve the benefit and risk ratio? Given the high mortality from breast cancer in African-American women, specific research attention should be given to the potential benefits and risks for African-American women in their forties. More information is also needed on the effectiveness of mammography in other racial or ethnic groups, including Native Americans, Hispanics, and Asians.
8. Is there a relationship between known risk factors for breast cancer incidence and the effectiveness of mammography?
9. Does the effectiveness of mammography differ between premenopausal and postmenopausal women?
10. How does estrogen replacement therapy affect the sensitivity and specificity of mammography?
11. Is the risk of radiation-induced breast cancer from mammography increased in women with a genetic susceptibility to breast cancer?
12. Are there new modalities or approaches to screening that would result in lower false-positive rates and increased sensitivity, and would thus lead to fewer diagnostic procedures?
13. Would increased education and an informed consent process reduce mammogram-related anxiety? Would it improve undesirable consequences of false-negative or false-positive examinations?
14. Is there a difference in the biologic behavior of cancers that cannot be detected mammographically? Does this affect clinical prognosis and response to treatment?
15. Is there any evidence that radiation-induced breast cancers have different characteristics, including biologic behavior?
16. Does low-dose radiation affect the biologic behavior of existing cancers?
17. Can a registry be established to combine raw data from all RCTs to quantify the benefit of mammography and relate it to age and other relevant characteristics? Can such a registry

be established in a way that it could rapidly incorporate newly available data and facilitate ongoing analyses?

18. Can practical and clear patient education materials be developed to facilitate a woman's decision regarding mammography?

Conclusions

Mammography has been shown to effectively reduce breast cancer mortality in women ages 50-69. Currently available evidence from RCTs indicates that for women ages 40-49, during the first 7-10 years following initiation of screening, breast cancer mortality is no lower in women who were assigned to screening than in controls. Summary data indicate a 16 percent reduction in breast cancer mortality after about 10 years, with confidence intervals of 2-28 percent. However, although some studies find lower mortality from breast cancer in screened women after 10 years, others do not. A lower mortality could result from the original screening or from other factors, such as CBE or mammography offered to the women after age 50.

This issue is further complicated by the charge to the panel to focus on a broad age range—40-49 years. The rationale for the charge was that evidence for recommending mammography is strong for women ages 50 and above, but not as clear for 40- to 49-year-old women. It should be pointed out that of all the studies reviewed, only one was originally designed specifically to evaluate mammography in the 40- to 49-year-old age group. However, age is a continuum and biologically there is no abrupt change at age 50. Indeed, a 49-year-old woman is probably more similar to a 50-year-old woman than she is to a 40-year-old. Unfortunately, there are no data upon which to base recommendations for narrower age ranges. The panel concludes that presently available evidence does not warrant a universal recommendation for mammography screening of women ages 40-49. This conclusion does not preclude the possibility that older women in this age group might have a different balance of benefit and risk than do younger women. Data to support this possibility, however, are not presently available. The effects of different ages at menopause also remain to be explored.

The potential benefits of mammography for women in their forties include earlier diagnosis and the option to choose breast-conserving therapy. These benefits must be weighed against the risks or potential risks, including those associated with false-positive tests: further diagnostie tests that may be invasive, anxiety, and inconvenience, as well as potential risk from mammographic radiation. In addition, the impact of false reassurance given to women with false-negative screens must be considered, given the lower sensitivity of mammography in women in their forties compared with women in their fifties. Professional and public education as well as disclaimers on mammography reports have increased awareness of false negatives in women with clinical symptoms such as a palpable lump. Similarly, those recommending mammographic screening of asymptomatic women in this age group must also remind women and their physicians to perform regular CBEs and to evaluate new symptoms promptly.

Every decision to utilize or not utilize a health-related service involves weighing available scientific evidence regarding benefits and risks against personal values and prior experiences. There are different levels of decision making, and the decision-making process will differ at each one. One level is characterized by the personal question, Would you have this done for yourself or for someone in your immediate family? When the available scientific evidenee is equivocal and incomplete, a person's decision to act or not act will be significantly influenced by personal or family experience with the disease and by one's capacity to deal with risk and uncertainty. Another level of deeision making is ''interpersonal,'' as when a physician decides to recommend a treatment to his or her patients. Such a decision is generally based more on the strength of the scientific evidence, but the physician's recommendations may also be colored by prior experience, both personally and with other patients, as well as by his or her assessment of the patient for whom the recommendation will be made. Finally, there is the large-scale level of decision making, such as when health officials decide to make across-the-board recommendations to a population, a decision that has far-reaching implications and that must be based to a much greater extent on a rigorous examination of the available scientific evidence. Of all decision levels, this level requires the strongest evidence of high benefit and low risk, particularly in the case of screening mammography, where such recommendations would be made to a healthy population. Thus, in some cases, a physician might recommend mammography for a patient in her forties and might do so despite a belief that the evidence is not sufficiently strong to warrant across-the-board recommendations.

The panel concludes that the data currently available do not warrant a universal recommendation for mammography for all women in their forties. Each woman should decide for herself whether to undergo mammography. Her decision may be based not only on an objective analysis of the scientific evidence and consideration of her individual medical history, but also on how she perceives and weighs each potential risk and benefit, the values she places on each, and how she deals with uncertainty. However, it is not sufficient just to advise a woman to make her own decision about mammograms. Given both the importance and the complexity of the issues involved in assessing the evidence, a woman should have access to the best possible relevant information regarding both benefits and risks, presented in an understandable and usable form. Information should be developed for women in their forties regarding potential benefits and risks so that each woman can make the most appropriate decision. In addition, educational material to accompany this information should be prepared to lead women step by step through the appropriate use of this information. For women in their forties who choose to have mammography performed, the costs of the mammograms should be reimbursed by third-party payors or covered by health maintenance organizations so that financial impediments will not influence a woman's decision.

Many women will seek guidance from their physicians who may be primary care physicians or physicians in different specialties. A woman's health care provider must be equipped with sufficient information to facilitate her decision-making process. Therefore, educational material for physicians should be developed to assist them in providing the guidance and support needed by their patients who are making difficult decisions regarding mammography.

A system should be established for ongoing monitoring and

review of newly available information from research studies regarding benefits and risks of mammography for women in their forties. This will ensure timely formulation and implementation of any new policy recommendations that may become appropriate in the future.

Minority Report

We, the undersigned members of the panel, have different interpretations of and derive different conclusions from the available data. We state those differences below.

1) *Is there a reduction in mortality from breast cancer due to screening women ages 40-49 with mammography, with or without physical examination? How large is the benefit? How does this change with age?*

Results from the eight RCTs indicate a statistically significant 17 percent mortality reduction ($P = 0.05$) for women ages 40-49 at time of entry into the trials. Although this survival benefit is less, on a population basis, than the benefit for women in older decades, it is nevertheless substantial. Furthermore, the potential biases in the RCTs would act to underestimate this benefit.

2) *What are the risks associated with screening women ages 40-49 with mammography, with or without physical examination? How large are the risks? How do they change with age?*

Although there is a theoretical risk from radiation exposure, if it exists at all, it is very low. There is no measurable harm from the diagnostic radiation doses used for screening mammography.

The majority statement discusses potential harm from false-negative mammograms and the potential for adverse psychosocial consequences from abnormal mammograms, but there are no data to support or quantify these possibilities.

The majority statement suggests that detection of DCIS is a potential harm. However, it is important to remember that all breast epithelium is within the ductal system. Therefore, biologically, all invasive ductal and lobular cancers must begin as *in situ* lesions. We do not know which DCIS will become invasive cancer and which will not. All DCIS is classified as cancer and must be taken seriously. Hence, detecting *in situ* cancer is a goal of and therefore a benefit of screening mammography rather than a harm.

An important risk for consideration is false-positive mammograms. These occur at all ages, lead to additional studies, and may cause anxiety and inconvenience. They constitute a measurable risk about which all women should be informed. Reported false-positive rates in mammography vary widely. Many of the studies reporting such data do not include sufficient detail to determine whether these rates vary significantly with decade of age. However, from the available data, it is reasonable to conclude that the false-positive rates for women in the 40-49 age range are higher than for older women, but only slightly higher than for women ages 50-59. False-positive mammograms that lead to additional views or breast ultrasound are generally considered to be of little consequence. The more important group of false positives are those that lead to biopsies for benign disease. The estimate of 25 percent (two cancers per eight biopsies) given in the majority statement is reasonable to expect for women in the 40-49 age group.

3) *Are there other benefits? If so, what are they? How do they change with age?*

The majority statement states, "Additional benefits . . . *may* include earlier detection" (italics added). There are unequivocal data indicating that screening mammography in women ages 40-49 does result in earlier detection. This earlier detection is an important benefit apart from any survival benefit. Detection at an earlier stage allows women more choice in treatment options.

The majority statement states, "Increased detection of DCIS *may* prove beneficial if it leads to a subsequent decrease in the incidence of invasive cancer" (italics added). We believe the data do indicate that increased detection of DCIS leads to a subsequent decrease in the incidence of invasive cancer, and this is a highly desirable goal.

There are not sufficient reported data to quantify the difference in these benefits by age within the 40-49 age group. However, the incidence of DCIS is similar across age groups.

In conclusion, we believe that the majority statement understates the benefits of mammography for women ages 40-49 and overstates the potential risks. We believe the data show a statistically significant mortality reduction for women in their forties. We further believe the survival benefit and diagnosis at an earlier stage outweigh the potential risks.

There are no data to suggest that women are significantly harmed by having extra mammographic views or breast ultrasound. Furthermore, the false-positive biopsy rate for mammography is not different from the false-positive biopsy rate for clinical breast examination. Moreover, the false-positive biopsy rate for women ages 40-49 is only slightly higher than for women ages 50-59, an age range for which mammographic screening is widely recommended.

Given our current understanding of breast cancer, it is potentially dangerous to suggest that DCIS may not be clinically important in women ages 40-49 and could safely be left undetected until women are in their fifties. Questioning the benefits of mammography for women ages 40-49 may cause significant harm from delayed diagnosis.

A majority of the panel did not accept that a statistically significant mortality reduction exists for women in their forties and so they were unable to make a universal recommendation for screening in this age group. We believe there is a statistically significant mortality reduction. We therefore recommend screening all healthy women in their forties. If we believe a certain recommendation is right for a 45-year-old family member, we would (and do) make the same recommendation to 45-year-old patients who come for advice and for 45-year-old women in general. We would alter that recommendation only if there were characteristics of the individual that were relevant. We agree that women should know what data and value judgments we use to form our recommendations, and we support their right to disagree with or reject our advice.

In summary, after evaluating and considering the evidence, we believe that we should actively encourage routine screening mammography for women in their forties. We also believe that providing accurate information to women and their health care providers is essential to assist women in deciding whether to accept or reject that advice.

DANIEL C. SULLIVAN, M.D.
RUTHANN T. ZERN, M.D.

Consensus Development Panel

Leon Gordis, M.D.
Conference and Panel Chairperson
Professor
Department of Epidemiology
School of Hygiene and Public Health
Associate Dean for Admissions and Academic Affairs
School of Medicine
Johns Hopkins University
Baltimore, MD

Donald A. Berry, Ph.D.
Professor
Institute of Statistics and Decision Sciences and Cancer Center Biostatistics
Duke University
Durham, NC

Susan Y. Chu, Ph.D., M.P.H.
Associate Director
Center for Health Studies
Group Health Cooperative of Puget Sound
Seattle, WA

Laurie L. Fajardo, M.D.
Professor of Radiology and Vice Chair for Research
Department of Radiology
University of Virginia
Charlottesville, VA

David G. Hoel, Ph.D.
Professor and Chairman
Department of Biometry and Epidemiology
Medical University of South Carolina
Charleston, SC

Leslie R. Laufman, M.D.
Hematology Oncology Consultants
Columbus, OH

Constance A. Rufenbarger
Project Development
The Catherine Peachey Fund, Inc.
Warsaw, IN

Julia R. Scott, R.N.
President and CEO
National Black Women's Health Project, Inc.
Washington, DC

Daniel C. Sullivan, M.D.
Associate Professor of Radiology
University of Pennsylvania Medical Center
Philadelphia, PA

John H. Wasson, M.D., F.A.C.P.
Herman O. West Professor of Geriatrics
Center for the Aging
Dartmouth Medical School
Hanover, NH

Carolyn L. Westhoff, M.D., M.S.
Associate Professor
Obstetrics, Gynecology, and Public Health
Columbia University College of Physicians and Surgeons
New York, NY

Ruthann T. Zern, M.D., F.A.C.O.G.
Obstetrician/Gynecologist
Private Practice
St. Joseph's Hospital
Greater Baltimore Medical Center
Towson, MD

Speakers

Freda E. Alexander, M.A., Ph.D., M.S.C.
"Basic Designs of Randomized Clinical Trials of Screening"
Department of Public Health Sciences
University of Edinburgh Medical School
Edinburgh, Scotland

Ingvar Andersson, M.D., Ph.D.
"The Malmö Mammographic Screening Trial: Update on Results and a Harm-Benefit Analysis"
Department of Diagnostic Radiology
University Hospital of Malmö, MAS
Malmö, Sweden

Cornelia J. Baines, M.D., M.S.C., F.A.C.E.
"Mammography Versus Clinical Examination of the Breasts"
Associate Professor
Department of Preventive Medicine and Biostatistics
Faculty of Medieine
University of Toronto
Toronto, ON, Canada

Nils Bjurstam, M.D., Ph.D.
"The Gothenburg Breast Screening Trial: Results from 11 Years' Follow-up"
Radiology Clinic
Section of Mammography
NÄL Hospital
Trollhättan, Sweden

Zora Kramer Brown
"A Breast Cancer Survivor's Perspective"
Founder and Chairperson
Breast Cancer Resource Committee
Washington, DC

Blake Cady, M.D.
"Detection and Treatment Trends: A Clinical Experience"
Professor of Surgery
Harvard School of Medicine
Interim Chief of Surgery
Surgical Oncology Division
Department of Surgery
Beth Israel Deaconess Medical Center
Boston, MA

Helena R. Chang, M.D., Ph.D., F.A.C.S.
"Screening for Breast Cancer in Younger Women Ages 40-49"
Surgeon and Director of the Hybridoma Laboratory
Associate Professor of Surgery and Pathobiology Program
Roger Williams Medical Center
Brown University
Providence, RI

Brian Cox, M.D., Ph.D., F.A.F.P.H.M.
"Variation in the Effect of Breast Screening by Year of Follow-up"
Senior Research Fellow and Public Health Physician
Department of Preventive and Social Medicine
University of Otago Medical School
Dunedin, New Zealand

Harry J. de Koning, M.D., Ph.D.
"Quantitative Interpretation of Age-Specific Mortality Reductions from Trials by Microsimulation"
Assistant Professor of Public Health/Medical Technology Assessment
Department of Public Health
Erasmus University
Rotterdam, The Netherlands

Stephen W. Duffy, M.Sc.
"Markov Models for Breast Tumor Progression: Estimates from Empirical Screening Data and Implications for Screening"
Senior Scientist, MRC Biostatistics Unit
Institute of Public Health
University Forvie Site
Cambridge University
Cambridge, U.K.

Virginia L. Ernster, Ph.D.
"Increases in Ductal Carcinoma *In Situ* in Relation to Mammography: A Dilemma"
Professor and Vice Chair
Department of Epidemiology and Biostatistics
Associate Director of the Cancer Center
School of Medicine
University of California, San Francisco
San Francisco, CA

Stephen A. Feig, M.D.
"Radiation Risk"
Professor of Radiology
Jefferson Medical College
Director, Breast Imaging
Department of Radiology
Thomas Jefferson University Hospital
Philadelphia, PA

Suzanne W. Fletcher, M.D., M.Sc.
"Breast Cancer Screening Among Women in Their Forties: An Overview of the Issues"
Professor
Primary Care Division
Department of Ambulatory Care and Prevention
Harvard Medical School and Harvard Pilgrim Health Care
Department of Epidemiology
Harvard School of Public Health
Boston, MA

Jan Frisell, M.D.
"The Stockholm Mammographic Screening Trial: Risks and Benefits"
Assistant Professor
Department of Surgery and Oncology
Stockholm South Hospital
Stockholm, Sweden

Paul Glasziou, M.B.B.S., Ph.D.
"The Quality and Interpretation of Mammographic Screening Trials for Women Ages 40-49"
Senior Lecturer in Clinical Epidemiology
Department of Social and Preventive Medicine
University of Queensland Medical School
Herston, Queensland, Australia

Russell P. Harris, M.D., M.P.H.
"Variation of Benefits and Harms of Breast Cancer Screening with Age"
Assistant Professor of Medicine
Division of General Medicine and Clinical Epidemiology
Co-Director, Program on Health Promotion and Disease Prevention
University of North Carolina at Chapel Hill School of Medicine
Chapel Hill, NC

R. Edward Hendrick, Ph.D.
"Benefit of Mammography Screening in Women Ages 40-49: Current Evidence from Randomized Controlled Trials"
Associate Professor and Chief
Department of Radiology
University of Colorado Health Sciences Center
Denver, CO

Karla M. Kerlikowske, M.D.
"Efficacy of Screening Mammography: Relative and Absolute Benefit"
"Outcomes of Modern Screening Mammography"
Associate Director
Women Veterans Comprehensive Health Center
Veterans Affairs Medical Center
Assistant Professor
Department of Medicine, Epidemiology, and Biostatistics
University of California, San Francisco
San Francisco, CA

Daniel B. Kopans, M.D., F.A.C.R.
"Problems with the Randomized Controlled Trials of Screening and Inappropriate Analysis of Breast Cancer Data"
Associate Professor of Radiology
Harvard Medical School
Director of Breast Imaging
Department of Radiology
Massachusetts General Hospital
Wang Ambulatory Care Center
Boston, MA

Nancy C. Lee, M.D.
"Results from the National Breast and Cervical Cancer Early Detection Program, 1991-1995"
Associate Director for Science
Division of Cancer Prevention and Control
Centers for Disease Control and Prevention
Atlanta, GA

Michael N. Linver, M.D., F.A.C.R.
"Mammography Outcomes in a Practice Setting by Age: Prognostic Factors, Sensitivity, and Positive Biopsy Rate"
Director of Mammography
X-Ray Associates of New Mexico, PC
Clinical Associate Professor
Department of Radiology
University of New Mexico School of Medicine
Albuquerque, NM

Anthony B. Miller, M.B., F.R.C.P.
"The Canadian National Breast Screening Study: Update on Breast Cancer Mortality"
Director, National Breast Screening Study
Professor and Chairman
Department of Preventive Medicine and Biostatistics
University of Toronto
Toronto, ON, Canada

Maryann Napoli
"What Do Women Want To Know?"
Associate Director
Center for Medical Consumers
New York, NY

Lennarth Nyström
"Update of the Overview of the Swedish Randomized Trials on Breast Cancer Screening with Mammography"
Biostatistieian/Epidemiologist
Department of Epidemiology and Public Health
Umeå University
Umeå, Sweden

Eugenio Paci, M.D.
"Study Design II"
Epidemiology Unit
Center for the Study and Prevention of Cancer
Florence, Italy

Barbara K. Rimer, Dr.P.H.
"The Psychosocial Consequences of Mammography"
Director
Cancer Prevention, Detection, and Control Research
Cancer Control Department
Duke Comprehensive Cancer Center
Duke University Medical Center
Durham, NC

Sam Shapiro
"Periodic Screening for Breast Cancer: The Health Insurance Plan of Greater New York Randomized Controlled Trial"
Professor Emeritus
Health Policy and Management
School of Hygiene and Public Health
John Hopkins University
Baltimore, MD

Edward A. Sickles, M.D., F.A.C.R.
"Screening Outcomes: Clinical Experience with Service Screening Using Modern Mammography"
Professor of Radiology
Chief, Breast Imaging Section
Department of Radiology
University of California Medical Center
University of California, San Francisco
San Francisco, CA

Robert A. Smith, Ph.D.
"Screening Fundamentals"
Senior Director
Department of Cancer Control
American Cancer Society
Atlanta, GA

László Tabár, M.D.
"Recent Results from the Swedish Two-County Trial: The Effects of Age, Histological Type, and Mode of Detection"
Associate Professor and Director
Department of Mammography
Falun Central Hospital
Falun, Sweden

Planning Committee

John K. Gohagan, Ph.D.
Chairperson
Branch Chief
Early Detection Branch
National Cancer Institute
National Institutes of Health
Bethesda, MD

Jeffrey S. Abrams, M.D.
Senior Investigator
Cancer Therapy Evaluation Program
Division of Cancer Treatment, Diagnosis, and Centers
National Cancer Institute
National Institutes of Health
Bethesda, MD

Anne R. Bavier, M.N., F.A.A.N.
Deputy Director
Office of Research on Women's Health
Office of the Director
National Institutes of Health
Bethesda, MD

Frank Bellino, Ph.D.
Health Scientist Administrator
Biology of Aging Program
National Institute on Aging
National Institutes of Health
Bethesda, MD

Jerry M. Elliott
Program Management and Analysis Officer
Office of Medical Applications of Research
National Institutes of Health
Bethesda, MD

Virginia L. Ernster, Ph.D.
Professor and Vice Chair
Department of Epidemiology and Biostatistics
Associate Director of the Cancer Center
School of Medicine
University of California, San Francisco
San Francisco, CA

John H. Ferguson, M.D.
Director
Office of Medical Applications of Research
National Institutes of Health
Bethesda, MD

Leslie G. Ford, M.D.
Associate Director
Early Detection and Community Oncology Program
Division of Cancer Prevention and Control
National Cancer Institute
National Institutes of Health
Bethesda, MD

Leon Gordis, M.D.
Conference and Panel Chairperson
Professor
Department of Epidemiology
School of Hygiene and Public Health
Associate Dean for Admissions and Academic Affairs
School of Medicine
Johns Hopkins University
Baltimore, MD

William H. Hall
Director of Communications
Office of Medical Applications of Research
National Institutes of Health
Bethesda, MD

Douglas B. Kamerow, M.D., M.P.H.
Director
Office of the Forum for Quality and Effectiveness in Health Care
Agency for Health Care Policy and Research
Rockville, MD

Daniel B. Kopans, M.D., F.A.C.R.
Associate Professor of Radiology
Harvard Medical School
Director of Breast Imaging
Department of Radiology
Massachusetts General Hospital
Wang Ambulatory Care Center
Boston, MA

Barnett S. Kramer, M.D., M.P.H.
Deputy Director
Division of Cancer Prevention and Control
National Cancer Institute
National Institutes of Health
Bethesda, MD

Amy S. Langer, M.B.A.
Executive Director
National Alliance of Breast Cancer Organizations
New York, NY

Elaine Lee
Program Analyst
Planning, Evaluation, and Analysis Branch
National Cancer Institute
National Institutes of Health
Bethesda, MD

Nancy Lee, M.D.
Associate Director for Science
Division of Cancer Prevention and Control
Centers for Disease Control and Prevention
Atlanta, GA

Carl M. Mansfield, M.D., D.Sc., F.A.C.R., F.A.C.N.M.
Associate Director
Radiation Research Program
Division of Cancer Treatment, Diagnosis, and Centers
National Cancer Institute
National Institutes of Health
Bethesda, MD

Anthony B. Miller, M.B., F.R.C.P.
Director, National Breast Screening Study
Professor and Chairman
Department of Preventive Medicine and Biostatistics
University of Toronto
Toronto, ON, Canada

Sue Moss, Ph.D.
Cancer Screening and Evaluation Unit
Section of Epidemiology
Institute of Cancer Research
Sutton, Surrey, U.K.

Nancy Nelson
Science Writer
National Cancer Institute
National Institutes of Health
Bethesda, MD

Gillian Newstead, M.D.
Director of Breast Imaging
Breast Imaging Center
New York University
New York, NY

Cherie Nichols, M.B.A.
Branch Chief
Planning, Evaluation, and Analysis Branch
National Cancer Institute
National Institutes of Health
Bethesda, MD

Vivian W. Pinn, M.D.
Director
Office of Research on Women's Health
Office of the Director
National Institutes of Health
Bethesda, MD

Alan Rabson, M.D.
Deputy Director
National Cancer Institute
National Institutes of Health
Bethesda, MD

Barbara K. Rimer, Dr.P.H.
Director
Cancer Prevention, Detection, and Control Research
Cancer Control Department
Duke Comprehensive Cancer Center
Duke University Medical Center
Durham, NC

Sam Shapiro
Professor Emeritus
Health Policy and Management
School of Hygiene and Public Health
Johns Hopkins University
Baltimore, MD

Edward A. Sickles, M.D., F.A.C.R.
Professor of Radiology
Chief, Breast Imaging Section
Department of Radiology
University of California Medical Center
San Francisco, CA

Robert A. Smith, Ph.D.
Senior Director
Department of Cancer Detection and Treatment
American Cancer Society
Atlanta, GA

Edward Sondik, Ph.D.
Director
National Center for Health Statistics
Centers for Disease Control and Prevention
Hyattsville, MD

László Tabár, M.D.
Associate Professor and Director
Department of Mammography
Falun Central Hospital
Falun, Sweden

Robert E. Tarone, Ph.D.
Mathematical Statistician
Division of Cancer Epidemiology and Genetics
National Cancer Institute
National Institutes of Health
Bethesda, MD

Rosemary Yancik, Ph.D.
Chief, Cancer Section
Geriatrics Program
National Institute on Aging
National Institutes of Health
Bethesda, MD

Conference Sponsors

National Cancer Institute
Richard D. Klausner, M.D.
Director

Office of Medical Applications of Research
National Institutes of Health
John H. Ferguson, M.D.
Director

Conference Cosponsors

National Institute on Aging
Richard J. Hodes, M.D.
Director

Office of Research on Women's Health
National Institutes of Health
Vivian W. Pinn, M.D.
Director

Centers for Disease Control and Prevention
David Satcher, M.D., Ph.D.
Director

Bibliography

The following references were provided by the speakers listed above and were neither reviewed nor approved by the panel.

Alexander FE, Anderson TJ, Brown HK, Forrest AP, Hepburn W, Kirkpatrick AE, et al. The Edinburgh randomised trial of breast cancer screening: results after 10 years of follow-up. Br J Cancer 1994;70:542-8.

American College of Radiology. Breast imaging reporting and data system. 2nd ed. Reston (VA): American College of Radiology, 1995.

Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. BMJ 1988;297:943-8.

Baines CJ. Women and breast cancer: is it really possible for the public to be well informed? [editorial]. Can Med Assoc J 1992;146:2147-8.

Baines CJ, Miller AB, Bassett AA. Physical examination. Its role as a single screening modality in the Canadian National Breast Screening Study. Cancer 1989;63:1816-22.

Beemsterboer PM, Warmerdam PG, Boer R, de Koning HJ. Radiation risk of mammography related to benefit in screening programmes: a favourable balance? Unpublished manuscript.

Bjurstam N, Bjorneld L. Mammography screening in women aged 40-49 years at entry: results of the randomized, controlled trial in Gothenburg, Sweden. Presented at 26th National Conference on Breast Cancer, 1994 May 10; Palm Desert (CA).

Black WC, Nease RF Jr, Tosteson AN. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. J Natl Cancer Inst 1995;87:720-31.

Black WC, Welch HG. Advances in diagnostic imaging and overestimations of disease prevalence and the benefits of therapy. N Engl J Med 1993;328:1237-43.

Breslow NE, Day NE. Statistical methods in cancer research. Volume II—The design and analysis of cohort studies. IARC Sci Publ 1987;82:1-406.

Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR Am J Roentgenol 1995;165:1373-7.

Brown ML, Kessler LG, Rueter FG. Is the supply of mammography machines outstripping need and demand? An economic analysis. Ann Intern Med 1990; 113:547-52.

Bull AR, Campbell MJ. Assessment of the psychological impact of a breast screening programme. Br J Radiol 1991;64:510-5.

Byrne C, Smart CR, Cherk C, Hartmann WH. Survival advantage differences by age. Evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project. Cancer 1994;74(1 Suppl):301-10.

Cady B. Is axillary node dissection necessary in routine management of breast cancer? No. In: DeVita VT Jr, Hellman S, Rosenberg SA, editors. Important advances in oncology. Philadelphia: Lippincott-Raven, 1996:251-65.

Cady B, Stone MD, Schuler JG, Thakur R, Wanner MA, Lavin PT. The new era in breast cancer. Invasion, size, and nodal involvement dramatically decreasing as a result of mammographic screening. Arch Surg 1996;131:301-8.

Cady B, Stone MD, Wayne J. New therapeutic possibilities in primary invasive breast cancer. Ann Surg 1993;218:338-47.

Campbell HS, McBean M, Mandin H, Bryant H. Teaching medical students how to perform a clinical breast examination. Acad Med 1994;69:993-5.

Campbell HS, Fletcher SW, Pilgrim CA, Morgan TM, Lin S. Improving physicians' and nurses' clinical breast examination: a randomized controlled trial. Am J Prev Med 1991;7:1-8.

Chen HH, Duffy SW, Tabar L. A Markov chain method to estimate the tumor rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. The Statistician 1996;45:307-17.

Cole P, Morrison AS. Basic issues in population screening for cancer. J Natl Cancer Inst 1980;64:1263-72.

Committee and Collaborators, Falun Meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40-49, Falun, Sweden, March 1996. Int J Cancer. In press.

Cox B. Benefit of mammography screening in women ages 40-49 years. Current evidence from randomized controlled trials [letter]. Cancer 1996;78:572-3.

Curpen BN, Sickles EA, Sollitto RA, Ominsky SH, Galvin HB, Frankel SD. The comparative value of mammographic screening for women 40-49 years old versus women 50-64 years old. AJR Am J Roentgenol 1995;164:1099-103.

de Koning HJ. Current controversies in cancer. Is mass screening for breast cancer cost-effective? Eur J Cancer 1996;32A:1835-44.

de Koning HJ, Boer R, Warmerdam PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials. J Natl Cancer Inst 1995;87:1217-23.

Duffy SW, Chen HH. Parameters of screening practice and outcome variation among programmes. Proceedings of the Falun Conference, 1996.

Eckhardt S, Badellino F, Murphy GP. UICC meeting on breast-cancer screening in pre-menopausal women in developed countries. Geneva, 29 September-1 October 1993. Int J Cancer 1994;56:1-5.

Elwood JM, Cox B, Richardson AK. The effectiveness of breast cancer screening by mammography in younger women [published errata appear in Online J Curr Clin Trials 1993; Doc No. 34 and 1994; Doc No. 121]. Online J Curr Clin Trials 1993; Doc No. 32.

Ernster VL. Epidemiology and natural history of ductal carcinoma in situ. In: Silverstein MJ, editor. Ductal carcinoma in situ of the breast: a diagnostic and therapeutic dilemma. Baltimore: Williams & Wilkins. In press.

Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996;275:913-8.

Feig SA. Estimation of currently attainable benefit from mammographic screening of women aged 40-49 years. Cancer 1995;75:2412-9.

Feig SA. Assessment of radiation risk from screening mammography [editorial]. Cancer 1996;77:818-22.

Feig SA. Determination of mammographic screening intervals with surrogate measures for women aged 40-49 years [editorial]. Radiology 1994;193: 311-4.

Feig SA, Dodd GD, Hendrick RE. Mammography risks and benefits. In: Radiation protection in medicine. Proceedings of the Twenty-eighth Annual Meeting of the National Council on Radiation Protection and Measurements, Proceedings No. 14. Bethesda (MD): National Council on Radiation Protection and Measurements, 1993:240-53.

Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85: 1644-56.

Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist LE, Somell A. Randomized study of mammography screening—preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18:49-56.

Frisell J, Glas U, Hellstrom L, Somell A. Randomized mammographic screening for breast cancer in Stockholm. Design, first rounds results and comparisons. Breast Cancer Res Treat 1986;8:45-54.

Frisell J, von Rosen A, Wiege M, Nilsson B, Goldman S. Interval cancer and survival in a randomized breast cancer screening trial in Stockholm. Breast Cancer Res Treat 1992;24:11-6.

Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. Med J Aust 1995;162:625-9.

Gustafsson L. QUEST. A program system for statistical and epidemiological data analysis. Umeå, Sweden: Umeå University, January 1990.

Hakama M. Breast cancer screening with mammography [letter]. Lancet 1993; 341:1531.

Harris R, Leininger L. Clinical strategies for breast cancer screening: weighing and using the evidence. Ann Intern Med 1995;122:539-47.

Hellman S. Karnofsky Memorial Lecture. Natural history of small breast cancers. J Clin Oncol 1994;12:2229-34.

Henson RM, Wyatt SW, Lee NC. The National Breast and Cervical Cancer Early Detection Program: a comprehensive public health response to two major health issues for women. J Public Health Management Practice 1996;2:36-47.

House Committee on Government Operations. Misused science: the National Cancer Institute's elimination of mammography guidelines for women in their forties. Union Calendar No. 480. House Report 103-863. October 20, 1994.

Hutchison GB, Shapiro S. Lead time gained by diagnostic screening for breast cancer. J Natl Cancer Inst 1968;41:665-81.

Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444-50.

Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276:33-8.

Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149-54.

Kopans DB. Mammography screening and the controversy concerning women aged 40 to 49. Rad Clin North Am 1995;33:1273-90.

Kopans DB, Feig SA. The Canadian National Breast Screening Study: a critical review. AJR Am J Roentgenol 1993;161:755-60.

Kopans DB, Halpern E, Hulka CA. Statistical power in breast cancer screening trials and mortality reduction among women 40-49 years of age with particular emphasis on the National Breast Screening Study of Canada. Cancer 1994;74: 1196-203.

Kopans DB, Moore RH, McCarthy KA, Hall DA, Hulka CA, Whitman GJ, et al. The positive predictive value of breast biopsy performed as a result of mammography: there is no abrupt change at age 50 years. Radiology 1996;200:357-60.

Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled Clin Trials 1981;2:93-113.

Larsson LG, Nystrom L, Wall S, Rutqvist L, Andersson I, Bjurstam N, et al. The Swedish randomised mammography screening trials: analysis of their effect on the breast cancer related excess mortality. J Med Screen 1996;3: 129-32.

Lein BC, Alex WR, Zebey DM, Pezzi CM. Results of needle localized breast biopsy in women under age 50. Am J Surg 1996;171:356-9.

Lerman C, Ross E, Boyce A, Gorchov P, McLaughlin R, Rimer BK, et al. The impact of mailed psychoeducational materials to women with abnormal mammograms. Am J Public Health 1992;82:729-30.

Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med 1991; 114:657-61.

Lidbrink E. Elfving J, Frisell J, Jonsson E. Neglected aspects of false positive findings of mammography in breast cancer screening: analysis of false positive cases from the Stockholm trial. BMJ 1996;312:273-6.

May DS, Jamison PM, Lee NC. Results from the National Breast and Cervical Cancer Early Detection Program, October 1 1991-September 30, 1993. MMWR Morb Mortal Wkly Rep 1994;43:530-4.

McCarthy BD, Yood MU, Boohaker EA, Ward RE, Rebner M, Johnson CC. Inadequate follow-up of abnormal mammograms. Am J Prev Med 1996;12:282-8.

Mettler FA, Upton AC, Kelsey CA, Ashby RN, Rosenberg RD, Linver MN. Benefits versus risks from mammography: a critical reassessment. Cancer 1996; 77:903-9.

Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years [published erratum appears in Can Med Assoc J 1993;148:718]. Can Med Assoc J 1992;147:1459-76.

Miller AB, Howe GR, Sherman GJ, Lindsay JP, Yaffe MJ, Dinner PJ, et al. Mortality from breast cancer after irradiation during fluoroscopic examinations in patients being treated for tuberculosis. N Engl J Med 1989;321:1285-9.

Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-4.

Morrison AS. Screening in chronic disease. New York: Oxford University Press, 1992.

Nielsen M, Thomsen JL, Primdahl S, Dyreborg U, Andersen JA. Breast cancer and atypia among young and middle-aged women: a study of 110 medicolegal autopsies. Br J Cancer 1987;56:814-9.

Nystrom L, Larsson LG, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, et al. Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials. A comparison between official statistics and validation by an endpoint committee. Acta Oncol 1995;34:145-52.

Nystrom L, Larsson LG, Wall S, Rutqvist LE, Andersson I, Bjurstam N, et al. The overview of the Swedish randomised mammography trials: total mortality pattern and the representivity of the study cohort. J Med Screen 1996;3:85-7.

Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341: 973-8.

Reese-Coulbourne J, National Breast Cancer Coalition, Washington, DC; Yaker A, executive director, Share, New York; Wiener B, director, Women's Cancer Resource Center, Minneapolis, MN; and the newsletters of Breast Cancer Action, San Francisco, and the Women's Community Cancer Project, Cambridge, MA.

Roberts MM, Alexander FE, Anderson TJ, Chetty U, Donnan PT, Forrest P, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241-6.

Roberts MM, Alexander FE, Anderson TJ, Forrest AP, Hepburn W, Huggins A, et al. The Edinburgh randomised trial of screening for breast cancer: description of method. Br J Cancer 1984;50:1-6.

Rutqvist LE. Summary of the characteristics of the Swedish trials. Proceedings of the Falun Conference, 1996.

Shapiro S, Venet W, Strax P, et al. Current results of the breast cancer screening randomized trial: the Health Insurance Plan (HIP) of greater New York study. In: Day N, Miller A, editors. Screening for breast cancer. Toronto: Hans Huber, 1988.

Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer: the Health Insurance Plan project and its sequelae, 1963-1986. Baltimore: Johns Hopkins University Press, 1988.

Shapiro S, Venet W, Strax P, et al. Selection, follow-up, and analysis in the Health Insurance Plan study: a randomized trial with breast cancer screening. In: Garfinkel L, Ochs O, Mushinski M, editors. Selection, follow-up, and analysis in prospective studies: a workshop. NIH Publ No. 85-2713; National Cancer Institute Monograph 67. Washington (DC): DHHS, PHS, 1985.

Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69:349-55.

Sickles EA, Kopans DB. Deficiencies in the analysis of breast cancer screening data [editorial]. J Natl Cancer Inst 1993;85:1621-4.

Skrabanek P. Breast cancer screening with mammography [letter]. Lancet 1993; 341:1531-2.

Smart CR, Hendrick RE, Rutledge JH 3rd, Smith RA. Benefit of mammography screening in women ages 40-49 years. Current evidence from randomized controlled trials. Cancer 1995;75:1619-26.

Tabar L, Duffy SW, Chen HH. Re: Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials. J Natl Cancer Inst 1996;88:52-5.

Tabar L, Duffy SW, Burhenne LW. New Swedish breast cancer detection results for women aged 40-49. Cancer 1993;72:1437-48.

Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Screening for breast cancer in women aged under 50: mode of detection, incidence, fatality, and histology. J Med Screen 1995;2:94-8.

Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187-210.

Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Tumour development, histology and grade of breast cancers: prognosis and progression. Int J Cancer 1996;66:413-9.

Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995;75:2507-17.

Tarone RE. The excess of patients with advanced breast cancer in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 1995;75:997-1003.

Thurfjell EL, Lindgren JA. Breast cancer survival rates with mammographic screening: similar favorable survival rates for women younger and those older than 50 years. Radiology 1996;201:421-6.

Thurfjell EL, Lindgren JA. Population-based mammography screening in Swedish clinical practice: prevalence and incidence screening in Uppsala county. Radiology 1994;193:351-7.

van Oortmarssen GJ, Habbema JD, van der Mass PJ, de Koning HJ, Collette HJ, Verbeek AL, et al. A model for breast cancer screening. Cancer 1990;66:1601-12.

Wald N, Chamberlain F, Hackshaw A. Report of the European Society for Mastology: Breast Cancer Screening Evaluation Committee (1993). Breast 1993; 2:209-16.

Walter SD, Day NE. Estimation of the duration of the pre-clinical state using screening data. Am J Epidemiol 1983;118:865-86.

About the NIH Consensus Development Program

NIH Consensus Development Conferences are convened to evaluate available scientific information and resolve safety and efficacy issues related to a biomedical technology. The resultant NIH Consensus Statements are intended to advance understanding of the technology or issue in question and to be useful to health professionals and the public.

NIH Consensus Statements are prepared by nonadvocate, non-Federal panels of experts, based on (1) presentations by investigators working in areas relevant to the consensus questions during a 2-day public session, (2) questions and statements from conference attendees during open discussion periods that are part of the public session, and (3) closed deliberations by the panel during the remainder of the second day and morning of the third. This statement is an independent report of the consensus panel and is not a policy statement of the NIH or the Federal Government.

Reference Information

For making bibliographic reference to this consensus statement, it is recommended that the following format be used, with or without source abbreviations, but without authorship attribution:

Breast Cancer Screening for Women Ages 40-49. NIH Consens Statement 1997 Jan 21-23; 15(1).

Publications Ordering Information

NIH Consensus Statements, NIH Technology Assessment Statements and related materials are available by writing to the NIH Consensus Program Information Center, P.O. Box 2577, Kensington, MD 20891; by calling toll free 1-888-NIH-CONSENSUS (888-644-2667); or by visiting the NIH Consensus Development Program home page on the World Wide Web at http://consensus.nih.gov

An Overview of the Breast Cancer Screening Controversy

*Daniel B. Kopans**

Randomized controlled studies show that screening mammograms are as important for women aged 40–49 as for women 50 years old and above. It was the improper use of retrospective, unplanned, sub-group analysis to advise women and their physicians that caused the controversy over mammograms for women under 50. Furthermore, arbitrarily grouping women into two groups leads to the incorrect conclusion that the age of 50 is a significant break point when it is not. The data demonstrates that none of the parameters of screening change abruptly at age 50. The recall rates (an abnormal mammogram) and the rate at which biopsies are recommended are virtually the same, regardless of age. Breast cancer is not a trivial problem for women in their forties. More than 30% of the years of life lost to breast cancer are from women diagnosed while in their forties. Because of changing demographics, in 1995 and 1996, there were actually more women diagnosed with breast cancer in their forties than for women in their fifties. The data clearly show that screening women for breast cancer, on an annual basis, beginning by age 40, can reduce the death rate by approximately 24%. It is important to separate medical and scientific analyses from the economic considerations. ''Society'' may decide that it is too expensive to screen women for breast cancer, but women should be provided with the scientific and medical information so that they can participate in the discussion of whether screening is ''worthwhile'' and decide whether or not to avail themselves of its benefit. The economics should not be used to influence the scientific and medical analysis of benefit. [Monogr Natl Cancer Inst 1997; 22:1–3]

There is now clear proof of benefit for screening women ages 40–49 for breast cancer. Not only have the randomized, controlled trials demonstrated a statistically significant mortality reduction of 18%, (*1*), but the Gothenburg trial has demonstrated a 44% mortality reduction that is statistically significant, by itself, and the Malmö trial has demonstrated a statistically significant reduction of 35% (presented to the NIH Consensus Development Conference, January 21–23, 1997). The data are now as strong as the results for women ages 50 and over, among whom only two trials are significant by themselves.

The benefit is even higher since the National Breast Screening Study (NBSS) of Canada should not be included in the analysis. Not only was it a trial of volunteers that differed from the 7 other trials that were trials by invitation, but the control group was screened by clinical breast examination unlike the unscreened controls in the other trials. Of greater concern is the fact that women with signs and symptoms of breast cancer were knowingly permitted to participate in the trial. This resulted in a major randomization problem (*2,3*) since the randomization was not blinded. All the women were first given a clinical breast examination and then were allocated to be screened, or to act as unscreened controls, based on open lists rather than blinded assignment. There were more women with lymph node positive cancers in the screened group than the controls. This has never equilibrated, as would be expected, suggesting an allocation imbalance. It resulted in 19 women with advanced breast cancer (4 or more positive nodes) being allocated to the screening arm, whereas there were only 5 women with advanced cancers allocated to the control arm. These are women who, not only could not be helped by screening, but who were likely to have died in the early years of follow-up. The explanation that the control women with breast cancer were treated in community hospitals and had fewer and less extensive axillary dissections than the screened women not only does not explain the imbalance, but it suggests a worrisome treatment asymmetry, as well, that could influence the results. The effort by MacMahon and Bailar to review the allocation process (*4*) was, unfortunately, inadequate since only a few centers were reviewed, and individuals who were involved in the allocation were never interviewed. The NBSS has yet to explain the excess of deaths that persist in the longer follow-up of the trial. Its results, by all estimates, make it a major outlier among the screening trials.

Why Has There Been a Controversy?

The randomized, controlled trials of breast cancer screening have actually, for many years, shown a statistically significant benefit for mammographic screening beginning by the age of 40. It was the inappropriate use of unplanned subgroup analysis that caused the confusion. The controversy over mammographic screening for women in their forties was not based on scientific analysis, but the incorrect use of data. With the exception of the NBSS, *none* of the RCTs were designed to evaluate women ages 40–49 as a separate group. *None* of the trials individually, or even collectively, had sufficient numbers of women in this decade of life to permit an expected benefit of 25% to be statistically significant in the early years of follow-up. In order to have an 80% power to demonstrate a 25% mortality reduction at five years (assuming a five-year survival of 75%), the trials would have had to involve almost 500,000 women split evenly into

**Affiliation of author:* Associate Professor of Radiology, Harvard Medical School, Cambridge, MA; Director of Breast Imaging, Massachusetts General Hospital, Boston.

Correspondence to: Daniel Kopans, M.D., Massachusetts General Hospital, Department of Radiology, Ambulatory Care Center, 15 Parkman Street, Second Floor, Room 219, Boston, MA 02114.

study and control groups (*5*). In addition to the fact that the trials were not designed to evaluate women ages 40–49 as a separate group (the screening intervals and techniques were not optimized) there were actually only 175,000 women under the age of 50 in all of the trials put together. Since it was mathematically impossible for an expected benefit of 25% to be statistically significant in the early years of follow-up, it was specious to suggest that there was no benefit when the benefits that did appear failed to reach significance (*6*). Advising women based on subgroup analysis of data from trials that lacked the statistical power to permit such analysis has been, at best, inappropriate, and the justification for this has never been provided. When analyzed as they were designed, however, the trials have, for many years, demonstrated a statistically significant benefit for screening beginning by the age of 40 (*7*). It is only the improper use of retrospective, unplanned, subgroup analysis to advise women and their physicians that caused the controversy.

Dichotomous Analysis Is Misleading

The confusion was compounded by reviews that purported to show abrupt changes in the parameters of screening occurring at the age of 50 (*8*). This was the result of data grouping that compared women ages 40–49 (as if they were a uniform group) to *all other women* ages 50 and over (as if they were a uniform group). This type of dichotomous grouping, making the age of 50 the point of analysis, leads to the fallacious interpretation and incorrect conclusion that the age of 50 is a significant break point when it is not. The data, in fact, when analyzed by smaller age groups, or individual age, demonstrate that the recall rates (an abnormal mammogram) are virtually the same, regardless of age and the rate at which biopsies are recommended is the same, regardless of age. The only thing that varies is the yield of cancer, and this changes gradually with increasing age, with no abrupt change at the age of 50, reflecting the prior probability of cancer in the population (*9*).

Despite the fact that the trials were not designed for sub-group analysis, with longer follow-up and more deaths, the trials now demonstrate statistically signifieant benefit, even when women ages 40–49 are analyzed separately. The most recent overview of the seven trials with similar design shows a 24% mortality reduction for women ages 40–49, that is significant. Even with the addition of the flawed NBSS data, the benefit is significant (*1*).

The Benefit Is Not Due to Women Reaching the Age of 50

The argument should be moot, but it has been suggested that this benefit is due to women reaching the age of 50 and screening suddenly becoming effective. Not only is this biologically not supportable, but RCT data cannot legitimately be analyzed by age at diagnosis. Age at diagnosis is a pseudovariable that is influenced by the intervention. Its use will, *a priori,* bias an analysis against cancers detected among younger women in the screened groups (*10*). RCT divide women into two groups. If the numbers involved are large enough, and the assignment is truly random, then every woman in the screened group will have a twin in the common group. For every woman in the screened group who develops a cancer there will be a woman in the control group whose cancer will behave in the same fashion. Using the age at diagnosis will bias the conclusions against the younger screened women. For example, assume that woman A (in the screened group) has her cancer detected when she is in her forties, and, as a consequenee, she will not die from breast cancer. Her "twin," patient B (in the control group), does not have her cancer diagnosed until she is in her fifties. If the age at diagnosis is used, the avoidance of death by "A" will not have any control group counterpart, and there will be no apparent mortality benefit for women screened in their forties. The death of woman "B" will be attributed to women over the age of 50. Thus, analyzing the data using the age at diagnosis will be misleading and will bias the results against screening the younger women. Nevertheless, even if the rules of RCT analysis are ignored and age at diagnosis is used, in the three trials that have performed such analyses, the benefit has been shown to be primarily for women whose cancers were diagnosed while they were still in their forties in the HIP trial (*11*), the Kopparberg trial (*12*), and in the Gothenburg trial (*1*).

The Benefit Is Actually Greater Than Indicated by the RCTs

What is often forgotten is that the RCTs underestimate the benefit of screening due to noncompliance and contamination. With the exception of the Canadian trial, which involved volunteers (a separate problem), the seven trials first randomized a population and then invited them to be screened. Women allocated to be screened who refused the invitation (noncompliance) are still counted as having been screened, and if they die of breast cancer their deaths are attributed to the screened group. Similarly, women who had mammograms on their own, outside of the screening program, and whose lives were saved as a result, are still counted as unscreened controls. The benefit of screening is likely higher than the trial results would indicate.

The "Harms" of Screening Do Not Change Suddenly at Age 50

Some analysts have raised the issue of "harms" from screening. These include anxiety from the process as well as biopsies that prove to be for a benign reason (termed unnecessary). Not only are these "harms" not equivalent to dying from breast cancer, but they are true for women at all ages, and do not change abruptly at the age of 50. As noted above, the recall rate for an abnormal mammogram is fairly constant across all ages, as is the "biopsy recommended" rate. The yield of breast cancer increases steadily with increasing age and merely reflects the prior probability of breast cancer in the population with no abrupt change at any age (*13*).

Breast Cancer Is Not a Trivial Problem for Women in Their Forties

Finally it has been suggested that breast cancer is not a major problem for women in their forties. In fact, more than 30% of the years of life lost to breast cancer are from women diagnosed while in their forties (*11*). Although the incidence of breast cancer increases steadily with increasing age, there are so many women in their forties, that, in 1995 and 1996, there were actually more women diagnosed with breast cancer in their forties than among women in fifties (*14*). It is also often forgotten that

many cancers that are diagnosed after the age of 50 have been growing for several years, and could have been diagnosed while the woman was in her forties.

A Delayed Benefit Does Not Mean No Benefit

Opponents have implied that, since the trials took longer for a benefit to appear among younger women than older women, that the benefit is not important. This is incorrect. To begin with, there is no biological reason to expect an immediate benefit. Given the parameters of the screening trials, a ''delayed'' benefit makes biological sense.

Most of the RCTs used a screening interval that was too long for younger women (two or more years between screens). Faster growing tumors were not interrupted. The benefit from interrupting the more moderate-growth cancers among the screened women cannot appear until the women in the control group succumb to their cancers. This is likely to not occur for five or more years after the cancers among the screened women were detected. Since most cancers are not detected in the first year of screening (the date from which the benefit is measured) and many women live for many years, even with breast cancer that will, ultimately, be lethal, the result is the appearance of a ''delayed'' benefit. Trials that screened at a shorter interval (Gothenburg and HIP) showed an earlier divergence of the mortality curves (years 5–7). Nevertheless, a ''delayed'' benefit does not lessen the value. As Feig has pointed out, a woman whose cancer is diagnosed at age 42 and consequently lives beyond age 52 derives as much if not more benefit than a woman whose cancer is found at age 55 such that she lives beyond age 60 (she had already lived beyond age 52).

The Determination of Medical Benefit Should Be Separated from Economics

It is important to separate the medical and scientific analysis from the economic considerations. ''Society'' may decide that it is too expensive to screen women for breast cancer, but women should be provided with the scientific and medical information, so that they can partieipate in the discussion of whether screening is ''worthwhile'' and decide whether or not to avail themselves of its benefit. The economics should not be used to influence the scientific and medical analysis of benefit.

Summary

The age of 50 has no biological significance, yet women and their physicians have been led to believe from data grouping and improper data analysis, that it represents a true threshold. There are no parameters of screening that change abruptly at age 50, or any other age. As with any test, there are false-negative examinations and false-positive examinations. Women at all ages should be provided with information concerning the ''risks'' and benefits of screening, so that they can make informed decisions.

The data elearly show that annually screening women for breast cancer, beginning by age 40, can reduee the death rate by approximately 24%. The benefit is likely even higher (*15*). Since there are no known ''risks'' that relate to an annual screening interval, women should know that the only reason to go to a longer interval between screens is economic. There is probably little or no radiation risk for women by the time they reach the age of 40 (*16*). Since the lead time for detecting cancer by mammography is approximately two years for younger women (it is not clear where ''younger'' ends and ''older'' begins) (*17,18*), screening at this interval, or longer, will not add much to the health care without screening. They should be screened at an interval that is less than two years (*19*). It may be possible to go to a longer interval among older women, since the lead time appears to be longer for them, but the age at which this can be done safely has not been determined. Since a 30% benefit has been shown for women over the age of 49 who were screened with intervals of almost three years, a much greater benefit will likely occur with more frequent screening.

References

(*1*) Tabar L, Larson LG, Andersson I, Duffy SW, Nystrom L, Rutqvist LE, et al. Breast cancer screening with mammography in women aged 40–49. Report of the organizing committee and collaborators, Falun Meeting, Falun, Sweden. March 21–22 1996. Int J Cancer 1996;68:693–9.

(*2*) Kopans DB, Feig SA. The Canadian National Breast Screening Study: a critical review. AJR AM J Roentgenol 1993;161:755–60.

(*3*) Tarone RE. The excess of patients with advanced breast cancers in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 1995;75:997–1003.

(*4*) Bailar JC, MacMahon B. Randomization in the Canadian National Breast Screening Study: a review for evidence of subversion. Can Med Assoc J 1997;156:193–9.

(*5*) Kopans DB, Halpern E, Hulka CA. Statistical power in breast cancer screening trials and mortality reduction among women 40–49 with particular emphasis on the National Breast Screening Study of Canada. Cancer 1994;74:1196–1203.

(*6*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.

(*7*) Shapiro S. Screening: assessment of current studies. Cancer 1994;74: 231–8.

(*8*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993:270:2444–50.

(*9*) Kopans DB, Moore RH, McCarthy KA, Hall DA, Hulka CA, Whitman GJ, et al. The positive predictive value of mammographically initiated breast biopsy: there is no abrupt change at age 50 years. Radiology 1996;200:357–60.

(*10*) Prorok PC, Hankey BF, Bundy BN. Concepts and problems in the evaluation of screening programs. J Chron Dis 1981;34;159–71.

(*11*) Shapiro S, Venet W, Strax P, Venet L. Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986. Baltimore (MD): Johns Hopkins University Press, 1988.

(*12*) Tabar L, Duffy SW, Chen HH. Re: quantitative interpretation of age-specific mortality reductions from the Swedish Breast Cancer Screening Trials. J Natl Cancer Inst 1996;88:52–3.

(*13*) Kopans DB, Moore RH, McCarthy KA, Hall DA, Hulka CA, Whitman GJ, et al. Biasing the interpretation of mammography screening data by age grouping: nothing changes abruptly at age 50. Presented at the Radiological Society of North America Meeting 1995, Chicago (IL).

(*14*) American Cancer Society statistics based on 1979–1993 SEER incidence rates and U.S. Census Projections provided by SEER. American Cancer Society, Atlanta (GA).

(*15*) Feig S. Estimation of currently attainable benefit from mammographic screening of women aged 40–49 years. Cancer 1995;75:2412–9.

(*16*) Mettler FA, Upton AC, Kelsey CA, Rosenberg RD, Linver MN. Benefits versus Risks from mammography: a critical assessment. Cancer 1996;77: 903–9.

(*17*) Moskowitz M. Breast cancer: age-specific growth rates and screening strategies. Radiology 1986;161:37–41.

(*18*) Tabar L, Faberberg G, Day NE, Holmberg L. What is the optimum interval between screening examinations? An analysis based on the latest results of the Swedish Two-county Breast Cancer Screening Trial. Br. J. Cancer 1987;55:547–51.

(*19*) Kerlikowske K, Grady D, Barclay J. Sickles EA, Ernster V. Likelihood ratios for modern screening mammography: risk of breast cancer based on age and mammographic interpretation. JAMA 1996;276:39–43.

Breast Cancer Screening Among Women in Their Forties: An Overview of the Issues

*Suzanne W. Fletcher**

This article summarizes the issues prompting a recent NIH Consensus Conference on mammography screening for women in their forties. To date, eight randomized controlled trials of breast cancer screening have been conducted, and a reduction in breast cancer mortality has emerged after 10 to 15 years of follow-up among women offered screening in their forties. No effect appears for at least eight years, and the reason for the delay, compared to that seen in women aged 50–69, is not clear. Two possibilities include cancer-stage shift due to screening in younger women and the aging of women into their fifties during the course of screening. Possible adverse effects of screening include radiation risk, although this is low, false-negative and false-positive screening tests, and overdiagnosis due to detection of ductal carcinoma *in situ* (DCIS). In order to make appropriate decisions regarding mammography, women need age-related information about both the benefits and potential risks of screening. [Monogr Natl Cancer Inst 1997;22:5–9]

Although 85% of breast cancers occur in women after they reach the age of 50, breast cancer is the number one cause of cancer death for American women aged 40–49. In 1993, it is estimated that 30,940 American women in this age group developed breast cancer and 4,843 died of it (Harras, A; personal communication). Each year, for every 100,000 women in their forties, 163 are diagnosed with breast cancer and 30 die of the disease (*1*).

Women in their forties need information to understand their risk for breast cancer. Data from SEER statistics indicate that for every 1,000 American women turning 40 years old, approximately 16 will develop breast cancer at some time before their fiftieth birthday (*1*). How many of these women will survive the cancer? SEER statistics show that nationally, 52% of women under 50 years of age who were diagnosed with invasive breast cancer in 1973 were still living 18 years later (*1*). Few, if any, of these women were likely screened. In the Health Insurance Plan (HIP) study, the only randomized controlled trial (RCT) with 18-year follow-up data, 58% of women in the group not offered screening survived to 18 years (*2*). With the advent of improved therapies over the past two decades, the percentage of women surviving breast cancer is improving (*3*). Thus, of the 16 women out of a thousand who will develop breast cancer in their forties, at least eight, and probably more, will survive the cancer regardless of screening. Therefore, breast cancer screening for women in their forties is primarily directed at the eight or fewer women in every 1,000 who might be saved by earlier detection of the cancer. If screening decreases mortality by as much as 25%, it would save one or two of the 16 women in a thousand who develop breast cancer in their forties.

Any potentially fatal illness striking persons in the prime of life is a terrible occurrence, but breast cancer is doubly so because it not only threatens a woman's life, but an emotionally and sexually important part of her body as well. Black and colleagues found that fear of breast cancer is so great that women in their forties overestimated their risk of dying of breast cancer 20-fold and their risk of developing breast cancer sixfold (*4*). With such a terrifying disease, it is important to find better ways to cure and prevent it.

What can screening, especially screening with mammography, contribute to the control of breast cancer in women in their forties? When considering screening for a particular medical disease, usually three questions are asked relating to the burden of disease, the characteristics of the screening test, and the effectiveness of early treatment (Table 1). In particular, it is important to examine the mortality benefits that accrue from the intervention, its adverse effects, and its costs. My task, then, is to present an overview of these issues as they pertain to breast cancer screening in women aged 40 to 49.

Breast Cancer Mortality Reduction

Most attention in breast cancer control has been directed towards determining the effect of screening on breast cancer mortality. Eight RCTs of mammography, with or without clinical breast examination, have been conducted in four countries: the HIP study from the United States; the Kopparberg, Östergötland, Malmö, Stockholm, and Gothenburg studies from Sweden; the Edinburgh study from the United Kingdom; and the National Breast Screening Study (NBSS-I) from Canada. At the National Cancer Institute International Workshop on Breast Cancer Screening in 1993, all seven trials published at that time found mortality reductions among women aged 50–69, two with statistically significant results (*5*). A meta-analysis presented at the Workshop found a statistically significant 34% reduction in breast cancer mortality after seven years of follow-up among women aged 50–69, with a relative risk ratio for screened to nonscreened women of 0.66 (95% confidence interval [CI]: 0.55–0.79) (*6*). However, the findings among younger women were less clear. The meta-analysis showed no effect at seven

**Affiliation of author:* Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, Boston, MA.

Correspondence to: Dr. Suzanne W. Fletcher, Department of Ambulatory Care and Prevention, Suite 200, 126 Brookline Ave., Boston, MA 02215.

Table 1. Criteria for deciding whether a medical condition should be included in periodic health examinations*

1. How great is the burden of suffering caused by the condition in terms of:			
	Death	Discomfort	
	Disease	Dissatisfaction	
	Disability	Destitution	
2. How good is the screening test, if one is to be performed, in terms of:			
	Sensitivity	Cost	Labeling Effects
	Specificity	Safety	
	Simplicity	Acceptability	
3. a. For primary prevention, how effective is the intervention?			
or			
b. For secondary prevention, if the condition is found, how effective is the ensuing treatment in terms of:			
	Efficacy		
	Patient Compliance		
	Early treatment being more effective than later treatment		

*Reprinted by permission from Fletcher R, Fletcher S, Wagner E. Clinical epidemiology—the essentials. Baltimore, Williams & Wilkins, 1996

years of follow-up, with a relative risk of 0.99 (95% CI: 0.74–1.32) or 1.08 (95% CI: 0.85–1.39), depending on whether or not results from the Canadian study were included.

A new overview of all five Swedish studies was also presented at the Workshop (*7*), and it showed a statistically insignificant 10% to 13% mortality reduction at 12 years of follow-up among women aged 40–49 (Fig. 1). This overview was more current than the meta-analysis because it included results from the Gothenburg trial that had not been previously published and because all cases in the other studies were re-reviewed, leading to some reassignment of subjects and outcomes. Nevertheless, the new Swedish analysis did not alter the conclusion from the meta-analysis that by seven years of follow-up, no beneficial effect is seen in younger women. With the data presented from the eight randomized trials, the Report of the International Workshop concluded, "For [women aged 40–49 years] it is clear that in the first 5–7 years after study entry, there is no reduction in mortality from breast cancer that can be attributed to screening. There is an uncertain and, if present, marginal reduction in mortality at about 10–12 years. Only one study (HIP) provides information on long-term effects beyond 12 years, and more information is needed."

In March 1996, an updated meta-analysis of the studies' results was reported in Falun, Sweden, for women aged 40–49 (Fig. 2). Five of the eight showed mortality reductions after 10–15 years of follow-up, and three showed no benefit. Pooled results demonstrated mortality reductions, with relative risk ratios of 0.77 (95% CI: 0.59–1.10), 0.76 (95% CI: 0.62–0.93), or 0.85 (95% CI: 0.71–1.01), depending on which trials were included.

As results of the trials continue to accrue, it appears that the time required to demonstrate beneficial screening effect, and the

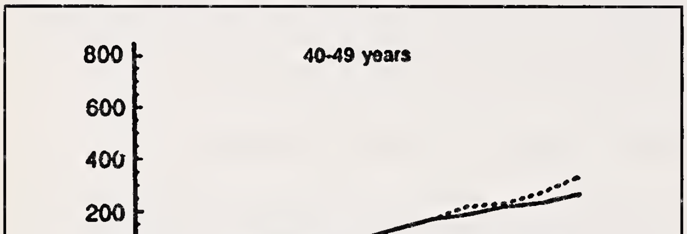

Fig. 1. Overview of Swedish randomized trials. Cumulative breast cancer mortality (per 1000) up to 12 years after randomization by age at randomization. Solid line = invited group and dotted line = control group. Adapted with permission from Nystrom et al. Breast cancer screening with mammography: overview of Swedish randomized trials. Lancet 1993;341:973–978.

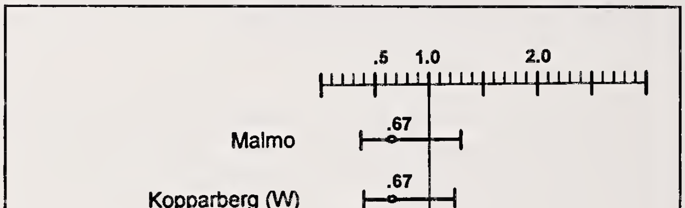

Fig. 2. Meta-analysis of randomized controlled trials of breast cancer screening for women in their forties, presented at Falun, Sweden; March, 1996. Relative risks are presented for each study, all Swedish trials, all population-based trials, and all trials. Adapted with permission from Breast cancer screening with mammography in women aged 40–49 years. Int J Cancer 1996;68:693–699. © Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc., 1996.

size of the effect, vary by age. Whereas mortality differences between screened and control groups began to emerge after only a few years of follow-up for women aged 50–69, the studies showed effects more slowly for women in their forties. In the combined Swedish studies (Fig. 1), mortality rates were similar in the invited and control groups during the first eight years of follow-up, after which a beneficial effect of screening began to emerge. This effect has continued to grow and is nearly statistically significant with three more years of follow-up (see Fig. 2, all Swedish trials). The same trend occurred in the HIP and Edinburgh studies. In most studies, breast cancer mortality reduction in younger women was less than in older women.

The cause of the time difference in effect by age is not yet clear. It has been suggested that screening picks up such early cancers in younger women that it takes longer for mortality reductions to occur (*8*). Indeed, randomized trials have shown that screening shifts diagnosis of breast cancer to earlier stages. Also, detection of ductal carcinomas is more common with the use of mammography. However, for stage shift to cause the time difference, screening would have to shift the stage of cancer differently according to age. Analysis of the Kopparberg trial showed shifts to earlier-stage cancers among women in both the forties and fifties (*9*). Data from all trials, however, should be examined to determine whether and to what degree stage shift could explain the delayed effect of screening in younger women.

The slower appearance of mortality reduction in younger women could also be partly due to ''age creep.'' Because screening studies occur over several years, some women who enter trials during their forties age into their fifties over the course of screening. It has been suggested that as women move into their fifties, when breast cancer screening is known to work, a benefit becomes apparent (*10,11*). Analyses of the RCTs are reported according to the age of women at entry into the trial, not their age at the time of diagnosis of breast cancer. This approach is necessary to preserve the comparability of the screened and control groups. Nevertheless, when there is the possibility that the effect of breast cancer screening varies by age, information about the age at diagnosis is needed. Most trials have not yet provided this information. The issue is especially important in the two trials in which only women 45 and older at entry were included.

Two groups have reported data about this issue. In the HIP study, 32% of cancers in women aged 40–49 at entry were detected after the women had turned 50 (*2*). Shapiro *et al.* demonstrated that screened women aged 45–49 at entry benefited when their cancers were diagnosed after age 50 but not when their cancers were diagnosed earlier. The numbers of women in each subgroup, however, were small. On the other hand, Tabar and Duffy did not find any effect of age creep in the Swedish two-county trial in which 36% of cancers were diagnosed after women in their forties turned 50 (*12*). The relative mortality was 0.95 (95% CI: 0.44–2.03) for women in whom breast cancer was diagnosed after they turned age 50 and 0.85 (95% CI: 0.49–1.45) for women in whom breast cancer was diagnosed before age 50. Information from the other trials is needed.

Ultimately, the degree to which age creep influences mortality effects for women in their forties will be best addressed by the results of the National Health Service Breast Screening Programme underway in the United Kingdom (*13*). In this large RCT, women aged 40 and 41 at entry are being screened annually for five years, which means that all breast cancers diagnosed will be in women who are still in their forties.

Why would a screening test for breast cancer have differential effects by age? Part of the explanation may be the lower accuracy (both sensitivity and specificity) of screening tests in younger women (*5*). Also, breast cancer growth rates may differ by age of the woman. Tabar *et al.* found that the mean sojourn time (time in the preclinical detectable state) was 1.25 years for women in their forties and 3.03 years for women in their fifties (*14*). Whether and how estrogen levels and menopause, rather than age per se, influence effectiveness of breast cancer screening remains unclear and needs to be determined. The question as to when to start breast cancer screening should not be arbitrarily linked to a particular decade of a woman's life.

It is important to determine the effect of screening in groups at high risk for breast cancer, especially in women aged 40–49. To date, no reports from the randomized trials have examined screening effects according to risk status of the participants.

Adverse Effects

Important possible adverse effects of breast cancer screening include radiation risk from mammography, adverse physical and psychosocial sequelae of false-positive and false-negative tests, and overdiagnosis.

Radiation risk from modern mammography appears to be very low. Feig and Ehrlich reviewed recent estimates for lifetime radiation risk to the breast and reported that a single mammogram exposure is estimated to cause between 2.9 and 8.8 excess breast cancers per million women screened during their forties (*15*). The estimates varied according to the age of the woman and the method of calculation, with more recent estimates being lower than older ones. If women in their forties are screened annually, radiation risk might cause between 29 and 88 additional breast cancers per million women screened for an entire decade.

Most previous work has analyzed the degree of accuracy of screening tests and the related problems of false-positive and false-negative results. The International Workshop Report summarized data from the randomized trials showing breast cancer screening in younger women is not as sensitive as in older women (*5*). Sensitivity by the detection method (ratio of screen-detected cancers to screen-detected plus interval cancers) among women aged 40–49 at study entry varied from 53% to 81% in the Stockholm, Swedish two-county, and Canada NBSS-I trials, while comparably defined sensitivity among women aged 50–59 varied from 73% to 88%. Thus, the risk of false-negative screening tests is greater in younger women.

Recent data suggest that false-positive mammograms—those requiring further evaluation to rule out cancer diagnosis—are a substantial problem in the United States. In a national survey of community mammography facilities conducted in 1992 and 1993, Brown *et al.* found that 11% (95% CI: 9%–13%) of screening mammograms were read as abnormal; 10.6% were false-positive readings (*16*). At a state-of-the-art program in northern California, Kerlikowske *et al.* reported that 6.3% of first-screen mammograms among women aged 40–49 were abnormal, and 5.9% proved to be false-positive readings (*17*).

Table 2 presents the percentage of women in each of these two studies who underwent additional procedures, including biopsies, following abnormal mammogram readings. On average, between one and two additional procedures were performed for each abnormal reading. Approximately 10% to 15% of women in each study underwent an invasive procedure. Lidbrink *et al.* have reported comparable results from the Stockholm trial (*18*).

The psychological effect of false-positive mammograms has been studied by Lerman *et al.* (*19*). They found that among women with high-suspicion mammograms that subsequently proved to be false-positive, three months later 47% reported being quite anxious about mammography, 41% were quite anxious about breast cancer, 26% reported that worry affected their mood, and 17% reported that it adversely affected their daily function. Women with low-suspicion mammograms reported less concern, but even among these women anxiety about mammography and breast cancer was relatively high (29% and 40% respectively, compared to 24% and 29% in women with normal mammograms). In a study from Norway, 18 months after screening mammography, 29% of 126 women with false-positive mammograms reported anxiety about breast cancer, compared to 13% of 152 randomly selected women with negative mammograms ($p = 0.001$) (*20*). In Britain, Ellman *et al.* found that 25% of women with normal mammograms, 30% of women with false-positive mammograms, and 35% of women with breast symptoms but with normal mammograms had General Health Questionnaire scores indicating probable psychiatric morbidity (*21*).

Because breast cancer screening is periodically repeated, it is important to determine the percentage of women who will experience a false-positive mammogram or clinical breast examination over an extended period of time. In the studies to date, it is clear that the percentage of false-positive mammograms decreases as women are rescreened. Nevertheless, it has been estimated that over a 10-year period of annual mammograms, as many as 30% of women in their forties will experience a false-positive mammogram or clinical breast examination (*22*). It is important to determine the actual number. Efforts to decrease the number of false-positive screening tests and their resultant adverse effects are also urgently needed.

As screening for breast cancer has increased, detection of ductal carcinomas *in situ* (DCISs) has risen—328% from 1983 to 1992 (*23*). An increasing percentage of breast cancers detected by screening are DCIS. In the northern Californian study discussed above, 43% of cancers detected among women in their forties were DCIS (*17*). Some experts are concerned that early lesions such as DCIS have led to overdiagnosis of breast cancer and is partly responsible for the recent increased incidence of breast cancer (*23*).

Table 2. Follow-up procedures for abnormal screening mammograms

	NSMF study* All ages, %	Northern CA study** 40–49 years, first screen %
Clinical breast examination	6.8	3.2
Repeat mammogram	37.0	—
Additional mammogram	41.2	56.1
Ultrasonography	20.0	10.9
Fine needle aspiration	3.0	5.8
Needle biopsy	2.9	—
Excisional biopsy	10.5	13.0
Needle localization	—	11.0

*National Survey of Mammography Facilities. Data from (*16*).
**Data from (*17*).

Finding breast cancer before any invasion, even microinvasion, has occurred should help save lives. However, there are a number of questions about DCIS. Pathologically, it appears difficult to diagnose. One study, for instance, asked six experienced pathologists to interpret 24 slides; there was complete agreement among the six in only two of the 10 cases in which at least one pathologist diagnosed DCIS (*24*). The prevalence and natural history of the condition are not clear and are important in determining if DCIS detection is leading to overdiagnosis. Finally, Ernster and colleagues have demonstrated that there is a wide range of treatment approaches for DCIS, not all of which may be appropriate (*23*). There is an urgent need for studies of all these issues.

In sum, progress has been made in better understanding the breast cancer mortality reduction that might occur with a screening program for women in their forties. Randomized trials have demonstrated that for approximately a decade, no benefit occurs, but after 10 to 15 years, a 15% to 25% mortality reduction appears. This translates into one or two women per 1,000 who might be saved. The reasons for the delay in mortality benefits and the degree to which these benefits could be achieved by beginning screening later remain to be determined. Also less clear are other benefits that might occur from earlier screening, such as more limited surgery or less debilitating adjuvant therapy. To achieve these benefits, however, substantial numbers of women will experience adverse effects, especially those caused by false-positive mammograms and possible overdiagnosis because of DCIS. Finally, costs of screening programs and the resultant procedures carried out because of abnormal readings cannot be ignored.

Women need information about all these issues. They rightly demand to be involved in an important decision about their lives and their bodies. Ultimately, it is the job of medical science to search for new and better ways to promote health, and along the way, to share with the public the very complicated facts as we understand them. Armed with facts, women can then apply their own set of values to cope with the important problem of breast cancer.

References

(*1*) Kosary CL, Ries LAG, Miller BA, Hankey BF, Harras A, Edwards BK. SEER Cancer Statistics Review, 1973–1992: Tables and graphs. Bethesda (MD): National Cancer Institute; 1995: DHHS Publ No. (NIH)96–2789.

(*2*) Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer: The Health Insurance Plan project and its sequelae; 1963–1986. Baltimore (MD): Johns Hopkins University Press, 1988.

(*3*) Chu KC, Tarone RE, Kessler LG, Ries LA, Hankey B, Miller BA, et al. Recent trends in US breast cancer incidence, survival and mortality rates. J Natl Cancer Inst 1996;88:1571–1579.

(*4*) Black WC, Nease RF Jr, Tosteson AN. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. J Natl Cancer Inst 1995;87:720–31.

(*5*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.

(*6*) Elwood JM, Cox B, Richardson AK. The effectiveness of breast cancer screening by mammography in younger women [published errata appear in Online J Curr Clin Trials 1993; Doc No. 34 and 1994; Doc No. 121]. Online J Curr Clin Trials 1993; Doc No. 32.

(7) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomized trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.
(8) Kopans DB. Mammography screening and the controversy concerning women aged 40 to 49. Radiol Clin North Am 1995;33:1273–90.
(9) Tabar L, Gad A, Holmberg L, Ljungquist U. Significant reduction in advanced breast cancer. Results of the first seven years of mammography screening in Kopparberg, Sweden. Diagn Imag Clin Med 1985;54: 158–164.
(10) Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.
(11) Fletcher SW. Why question screening mammography for women in their forties? Radiol Clin North Am 1995;33:1259–1271.
(12) Tabar L, Duffy SW, Chen HH. Re: Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials. J Natl Cancer Inst 1996;88:52–5.
(13) Moss SM, Michel M, Patnick J, Johns L, Blanks R, Chamberlain J. Results from the NHS breast screening programme 1990–1993. J Med Screen 1995;4:186–90.
(14) Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.
(15) Feig SA, Ehrlich SM. Estimation of radiation risk from screening mammography: recent trends and comparison with expected benefits. Radiology 1990;174:638–7.
(16) Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR Am J Roentgenol 1995;165: 1373–7.
(17) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.
(18) Lidbrink E, Elfving J, Frisell J, Jonsson E. Neglected aspects of false positive findings of mammography in breast cancer screening: analysis of false positive cases from the Stockholm trial. BMJ 1996;312:273–6.
(19) Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med 1991;114:657–61.
(20) Graham IT, Lund E, Slenker SE. Quality of life following a false positive mammogram. Br J Cancer 1990;62:1018–22.
(21) Ellman R, Angeli N, Christians A, Moss A, Chamberlain J, Macquire P. Psychiatric morbidity associated with screening for breast cancer. Br J Cancer 1989;60:781–4.
(22) Eddy DM. Screening for breast cancer. Ann Intern Med 1989;111:389–99.
(23) Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996; 275:913–8.
(24) Schnitt SJ, Connolly JL, Tavassoli FA, Fechner RE, Kempson RL, Gelman R, Page DL. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. Am J Surg Pathol. 1992;16:1133–43.

What Do Women Want to Know?

*Maryann Napoli**

In the early 1970s, before there was any scientific evidence to prove mammography's benefit to younger women, the American Cancer Society (ACS) and the National Cancer Institute (NCI) began to promote screening for all women over the age of 35. The ACS's message to the public was—and still is—"breast cancer is curable, if detected early enough." In 1985, mammography equipment companies and other businesses with vested interests in getting women to undergo screening began taking over the "public education" efforts with exaggerated claims, such as "a 91% cure rate." By the time the NCI withdrew its mammography screening recommendation to women in their forties, it was too late. Most women now overestimate their odds of developing breast cancer in their forties and overestimate what mammography can do for them. The recent NIH Consensus Conference Report on mammography screening could have a major impact by explaining that the overwhelming majority of breast cancers are unaffected by early detection, either because they are aggressive or slow growing. Women must be better informed about the risks of mammography screening, especially the uncertainties surrounding a diagnosis of ductal carcinoma *in situ.* [Monogr Natl Cancer Inst 1997; 22:11–13]

I would not presume to speak for all women today, but I draw upon the experiences of those who come to my organization, the Center for Medical Consumers. Breast cancer has brought hundreds of women to our medical library, which has been open to the public for over 20 years in order to promote informed decision making. I can also speak of what I learn from the growing number of breast cancer advocacy organizations (*1*).

And lastly, I speak for myself. As a medical writer, I have followed the literature on breast cancer. As a consumer advocate, I have followed the selling of mammography screening to women, ever since the early 1970s, when the Breast Cancer Detection and Demonstration Project (BCDDP) introduced the concept of mammography screening at 27 medical centers in the United States. About 280,000 women over age 35 took part in the BCDDP, which was sponsored by the American Cancer Society (ACS) and the National Cancer Institute (NCI).

My organization, the Center for Medical Consumers, is founded on the belief that people should be encouraged to base their medical treatment decisions on the published evidence. We also believe that screening decisions should be held to the highest standard of evidence because they affect healthy people.

When the NCI announced its 1993 decision to withdraw mammography screening recommendation for women in their forties, I believe that it made the correct judgment. But this didn't seem to change many opinions. Women had already been sold the idea that early detection of breast cancer at any age virtually guarantees cure. The two most common reactions I heard from women at that time were: "I'll still have mammograms just to play it safe"; and "What can we do to protect ourselves, if they take away mammography?" To many, it seemed inconceivable that finding a tumor early could be anything but beneficial. At the very least, many women reasoned, finding a breast cancer early would mean a less drastic treatment—a widespread misperception given the fact that breast-sparing treatment is appropriate for node-positive disease and tumors up to 4 cm (*2*). In a scenario I have observed many times, be it a public forum on breast cancer or a radio show, the speaker who points to the lack of scientific evidence to support mammography screening for younger women invariably triggers a response like this from a member of the audience: "How dare you say that mammography has no benefit to women in their forties; my breast cancer was discovered on a mammogram last year when I was 43. Now my life has been saved."

These reactions must be viewed against the backdrop of the "public education" surrounding the BCDDP and the more recent breast cancer awareness activities. The overly optimistic opinions surrounding mammography screening's value to women in their forties are the direct result of promoting a technology to the public before there was clear scientific evidence proving its benefit to younger women (*3,4*).

Soon after we opened our Center in 1977, I became aware of a book for women called *Early Detection: Breast Cancer is Curable* (*5*) by Dr. Philip Strax, the radiologist who co-authored the Health Insurance Plan (HIP) of Greater New York study (*6*). Before that study was over, Dr. Strax, then a spokesman for the ACS, established a prototype screening center in New York City called the Guttman Institute, which, in 1968, began offering mammography screening and breast exams to women over age 35 (*7*). With the best of intentions, I'm sure, mammography began to be promoted to women because the HIP study showed a short-term mortality reduction, despite the fact that mammography's role in that reduction was unclear (the study did not separate the effect of mammography from the clinical breast exam), and despite the fact that mammography's benefit to younger women was unproved (*8,9*). Over a decade later, the HIP results were re-examined by an independent committee of experts and found to have serious flaws. For example, half the cancers said to have been discovered by mammography alone were actually palpable and in no way clinically occult (*10*).

A woman's doctor may have the most influence in determining whether she will undergo mammography screening. Most doctors tell women in their forties to be screened because, in most cases, their professional organizations (*11*) advise them to

**Affiliation of author:* Center for Medical Consumers, New York, NY.

Correspondence to: Maryann Napoli, Associate Director, Center for Medical Consumers, 237 Thompson Street, New York, NY 10012–1090.

do so. But the most influential source of information for the lore surrounding mammography screening—for the overly optimistic expectations surrounding mammography—is the ACS. The ACS has a long history of overstating the case for early detection (*7*), of using five-year survival statistics to imply cure (*12*), of recommending screening tests before there is scientific evidence to prove safety and efficacy (*13*), and of not warning the public about the risks of screening. In the case of mammography there is the very real possibility of undergoing either an unnecessary mastectomy or unnecessary radiotherapy. Wider acceptance of mammography screening had led to a dramatic increase in the diagnosis of ductal carcinoma *in situ* (DCIS) (*20*). Many, perhaps most, of these microscopic lesions would never have progressed to invasive cancer even if left untreated, yet DCIS continues to be treated with mastectomy or radiotherapy in the majority of cases. In this regard, things haven't progressed much since the BCDDP. In 1977, the public learned about so-called microscopic cancers that caused 64 women to be misdiagnosed as having breast cancer during the BCDDP: 37 had undergone mastectomy (*4*). Quite a revelation. No one ever warns the public about finding a cancer so early that pathologists aren't sure that it's even cancer. And here we are, 20 years later, and pathologists are still trying to determine the natural history of the different subtypes of DCIS in order to avoid overtreatment (*14*).

Now there's a new generation of women in their forties who were too young at the time of those 1977 headlines to be concerned about mammography-related misdiagnoses. After all, breast cancer in that era was an older woman's disease. Women now in their forties have been "raised" on the public health message that "breast cancer is curable if found early enough." In other words, cure is simply a matter of finding breast cancer early. In other words, if you're dying of breast cancer, it's your fault because you didn't find it early enough.

Yet in 1980, I came across a New England Journal of Medicine review of all published breast cancer trials which found that 25%–35% of all women diagnosed and treated at Stage I developed metastasis anyway and died within 10 years of their mastectomies (*15*). This is just one of many contradictions I would find between the "public education" message to women and the published evidence.

In 1985, we saw the start of breast cancer awareness activities, initiated and largely sponsored by Zeneca, the manufacturer of tamoxifen. Now, it is the corporate ads like those of DuPont and General Electric (G.E.), makers of mammography-related equipment, that feature the same old misleading statistics. G.E.'s recent long-running television ad, for example, claimed "a remarkable 91% cure rate" for mammography. These corporate ads come cloaked in the aura of public service announcements (PSA). And frankly, in terms of half-truths, I don't find them to be any different than the real PSAs sponsored by the ACS or the American College of Radiology (*16*). The depiction of young women in these ads, the use of "one in eight" and "one in nine" statistics, the magazines and talk shows featuring personal stories of young breast cancer survivors all have contributed to the impression of breast cancer as a young woman's disease. Put this heightened awareness together with the exaggerated "public health" message—early detection equals cure—and you have a lot of women out there who think that a mammogram is the only thing that stands between them and imminent death from breast cancer.

Any honest public discussion of mammography screening's risk has been and still is discouraged. For example, when Dr. John C. Bailar, III, M.D., published his 1976 article stating, ". . . routine use of mammography in screening asymptomatic women may eventually take almost as may lives as it saves," (*17*) hostile radiologists called one UPI reporter, Patricia McCormick, to say she was causing breast cancer deaths by reporting Bailar's point of view (Personal communication, Patricia McCormick, who covered the BCDDP). Radiologists today take a similar stand against the reporting of mammography's risks because it might stop women from having mammography (*18*). This rather patronizing argument surfaces on those rare occasions when the topic of "overdiagnosis" makes it into the general media (*19*). (Notice how the medical word sanitizes the problem: physicians use the word "overdiagnosis" when they mean misdiagnosis, when they mean finding cancer that isn't there.) Mammography proponents invariably frame the debate in this manner: what's the harm of anxiety over an abnormal mammogram or a biopsy compared to death from breast cancer? Well, we don't know whether any deaths are prevented, and many women (including those over age 50) do not fully understand the third possibility associated with mammography screening: misdiagnosis of cancer. The overreading of atypical benign breast disease as carcinoma *in situ,* or of *in situ* disease as invasive cancer, has occurred in several major trials where pathologists would be expected to be more expert than those in the real world (*20*). I have met many a woman who has had a mastectomy for DCIS, who regards herself as a cancer survivor, who worries about recurrence like every other cancer patient, who believes her daughters are at high risk, and who has no idea of the uncertainties that surround her diagnosis or that evidence suggests that only some cases of DCIS will become invasive cancer. In the last few years, however, there has been a change. Most women today with a diagnosis of DCIS come to our Center knowing something about the controversies surrounding it. But the point is they hear it for the first time at diagnosis, not before they consent to screening in the first place.

In summary, women have received such one-sided and distorted information about early detection that most probably don't know what they should be asking about mammography screening in their forties. Women continue to hear to this day the same inflated message of Dr. Strax's book two decades ago: "Breast cancer is curable, if detected early enough" (*21*).

At this point, I would like to change the title of my speech to: "What Do Women Need to Know." A consensus pronouncement isn't enough unless you also educate the public about scientific evidence: about how mortality reduction proves the value of a screening test, not how many cancers it can find, and not the number of women in their forties who get cancer.

But there's always a part of me asking: Does anyone really care about scientific evidence? Do we accept clinical trial findings only when they support our well-entrenched ideas? I'm including doctors in my questions. Look how long it took surgeons and radiologists to let go of the Halsted-radical mastectomy, the modified radical mastectomy, and routine radiotherapy after modified or total mastectomy—just to cite a few examples.

When the National Breast Screening Study of Canada was

published, its design and mammographic techniques were attacked by American radiologists (*22*). Few women had the time or the skills to make an in-depth assessment of their arguments. The suggestion that Canadian mammography techniques were inferior to ours, however, seemed plausible to many women. But I found the "mammography has improved" argument troubling. Does this mean that medical technologies should never be subjected to controlled trials because the findings will always be obsolete by the time they are published? If mammography techniques have improved so much, why were the greatest mortality reductions shown for the two earliest trials (*23*)?

Nearly 30 years of promoting mammography screening have passed, and its proven success in reducing breast cancer mortality in older women has yet to be reflected in the nation's cancer statistics. Given the massive amount of resources poured into the study and promotion of this screening test, the return has been modest, at best. It is time to give priority to etiology. Little is known about how to prevent breast cancer. And I'm not talking here about giving a drug like tamoxifen to healthy women to see if it can prevent more caneers than it causes.

Over the last few years, I've noticed a change in thinking about mammography screening among the breast cancer survivor/activists. Traditionally, cancer survivors become evangelists for screening, but I've detected less enthusiasm of late (*24,25*). Every breast cancer activist I know is a woman diagnosed in her forties. These women know firsthand about mammography's other problem: its high false-negative rate for younger women. I have contacted several advocacy organizations and heard variations on this theme: "We'll continue to have mammograms, but researchers must find better ways to detect early breast cancers because mammography does not help most women. We need to know more about what causes breast cancer." Mammography may be the best detection tool we have, as the PSAs constantly remind women, but it's just not good enough.

Some activists are highly critical of the excessive focus on genetic research. They want more funding directed to a better understanding of carcinogens in the air, water, and food. Some challenge the NCI's focus on individual susceptibility rather than on social responsibility (*26*).

In closing, I want to address the new evidence from Sweden showing a reduction in breast cancer mortality. For nearly a year, radiologists have been portraying this finding to the public as the proof that now ends the controversy (*27,28*). As someone who listens to how people receive statistical information, I would urge the panel to give careful consideration to the layman's explanation of this new finding. What, for example, does the reduction in "subsequent" mortality actually mean? (The public never hears that qualifying word.) Is this finding an argument for starting screening at age 40, or for delaying it until age 50? How does a woman weigh the 16% reduction in subsequent morality against her odds of misdiagnosis? Does this new finding mean that everyone who undergoes mammography screening can reduce her personal odds of dying of breast cancer by 16% (which is how most people interpret such a statistic)? Or, is it fairer to put it this way: mammography screening may result in a prolonged life for 16% of women with breast cancers? The majority of women whose eancers are found on a mammogram, however, will be unaffected by early detection, either because they have an aggressive, fast-growing caneer or because the tumor is so slow growing the women would enjoy long-term survival whether it was found early on a mammogram or later, once a symptom appeared (*29*). Some women will be falsely assured that they are cancer-free.

Here is where the Consensus Panel could have the greatest impact—by offering a full and honest explanation of statistics, by educating women and their doctors about what mammography can and cannot do, and by bringing to this topic a large dose of reality.

References

(*1*) National Breast Cancer Coalition, Washington (DC); SHARE, New York City; Women's Cancer Resource Center, Minneapolis; Breast Cancer Action, San Francisco; Women's Community Cancer Project, Cambridge (MA); Action for Cancer Prevention Campaign, New York (NY).

(*2*) Fisher B, et al. Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer. N Engl J Med 1995;333:1456–61.

(*3*) Carbone PP. A lesson from the mammography issue. Ann Intern Med 1978;88:5,703–4.

(*4*) Final word on disputed mastectomies. Science 1978;202:728.

(*5*) Strax P. Early Detection: Breast Cancer is Curable. New York: Harper & Row, 1974.

(*6*) Shapiro S, Strax P, Venet L. Periodic breast cancer screening in reducing mortality from breast cancer. JAMA 1971;215:1777–85.

(*7*) Brody JE, Holleb, AL. You can fight cancer and win. New York: Quadrange/New York Times Book Co. 1974:65–9.

(*8*) Kunz J. Mammography dispute continues to simmer. JAMA 1977;238: 1999–2006.

(*9*) Greenberg DS. X-ray mammography—background to a decision. N Engl J Med 1976;295:13:739–40.

(*10*) Greenberg DS. X-ray mammography: silent treatment for a troublesome report. N Engl J Med 1977;296:17,1015–6.

(*11*) Screening for breast cancer. In: Guide to clinical preventive services. DHHS, PHS, DPHP, 1996:80–1.

(*12*) American Cancer Society. Facts on breast cancer, 1978. Breast cancer: questions & answers [pamphlet]. 1994.

(*13*) American Cancer Society report on the cancer-related health checkup. CA Cancer J Clin 1980;30:4,194–240.

(*14*) Schnitt SJ, Harris JR, Smith BL. Developing a prognostic index for ductal carcinoma in situ of the breast: are we there yet? Cancer 1996;77:11, 2189–92.

(*15*) Henderson IC, Canellos GP. Cancer of the breast: the past decade, Parts I and II. N Engl J Med 1980;302:78–90.

(*16*) Don't mess with your life. Have a mammogram. Public health message from the American College of Radiology-accredited clinics. The New York Times 1993 October 3.

(*17*) Bailar JC. Mammography: a contrary view. Ann Intern Med 1976;84:77–84.

(*18*) Kopans DB. Letter. Detection and treatment of ductal carcinoma in situ. JAMA 1996;276:869.

(*19*) Kolata G. Mammograms before 50? A hung jury. The New York Times 1993 November 14.

(*20*) Ketcham AS, Moffat FL. Vexed surgeons, perplexed patients, and breast cancer which may not be cancer. Cancer 1990;65:387–93.

(*21*) Anthony M. Office of Women's Health, Food and Drug Administration, speaking at "Update on breast cancer: federal agency overview," held at the Association of the Bar of New York City, 1996 October 10.

(*22*) Allison M. Mammography trial comes under fire. Science 1992;256: 1128–30.

(*23*) Wright CJ, Mueller CB. Screening mammography and public health policy: the need for perspective. Lancet 1995;346:29–32.

(*24*) Visco FM. National Breast Cancer Coalition letter to its membership. November 1996.

(*25*) Peterson N. Mammography under fire. Breast Cancer Action Newsletter, October/November 1996.

(*26*) Brenner B. Whose cancer institute is it anyway? Breast Cancer Action Newsletter, October/November 1996.

(*27*) American College of Radiology. Press release: Swedish study supports U.S. groups' position calling for mammography screening for women aged 44–49. 1996 March 21.

(*28*) Kopans DB. Don't let politics sway mammogram debate [letter]. The New York Times 1996 December 13.

(*29*) Lee JM. Screening and informed consent. N Engl J Med 1993;328:6, 438–40.

Screening Fundamentals

*Robert A. Smith**

While researchers have established the value of screening for breast cancer with mammography, with and without clinical breast examination, age-specific analyses have led to differing opinions regarding the ages and the intervals that breast cancer screening should begin. This article, therefore, provides a detailed, age-specific evaluation of mammography screening by assessing the severity of breast cancer, the effectiveness of earlier versus later treatment, and the accuracy and reliability of mammography. Data from previous randomized trials and other sources are used to evaluate these criteria. The results indicate that screening programs must have high levels of participation, achieve acceptable sensitivity (85%) and specificity (90%), adopt age-specific screening intervals, and consider how disease stage influences diagnosis. In addition, as others have noted, the following benchmarks can be used to evaluate screening programs: (1) more than 50% of screen-detected cancers should be smaller than 15 mm; (2) 30% or more of grade 3 cancers detected on screening should be less than 15 mm; and (3) more than 70% of cancers detected on screening should be node negative. [Monogr Natl Cancer Inst 1997;22:15–19]

As a disease control strategy and policy, the goal of breast cancer screening is to reduce morbidity and mortality by distinguishing those individuals in an asymptomatic population that are likely and not likely to have breast cancer (*1*). The emphasis on likelihood is important and inherent in the concept of screening. A person identified by a screening test as likely to have a disease is then referred for further diagnostic testing to determine whether he or she does in fact have the disease and therefore needs treatment. The emphasis on likelihood also is important because screening tests and programs have inherent limitations according to the criteria that will be described below; thus, while the majority of screening test interpretations are correct, inevitably some individuals will be incorrectly identified as possibly having the disease (a "false positive"), and screening will fail to identify some who do have the disease (a "false negative"). The advantage of screening an asymptomatic population is that the test can identify preclinical disease with sufficient *lead time*—that is, the time before the expected onset of symptoms—to potentially alter the natural, and more adverse, course of disease.

In order to be an effective disease control strategy, a screening program should meet fundamental criteria in three areas: 1) characteristics of the disease; 2) the effectiveness of earlier versus later treatment; and 3) characteristics of the screening test—specifically, its accuracy and reliability, but also its costs and acceptability to the target population (*2*). It would be ideal if there were conventional benchmarks for these criteria, either alone or considered together, but this is not the case. Further, decisions about screening are more easily reached if the evidence for the effectiveness of earlier versus later treatment, or test performance, derives from well-designed randomized clinical trials, since observational studies are subject to well-known biases that complicate the interpretation of end results (*2*). When such evidence is lacking, decision makers are confronted with two alternatives: await data from a well-designed randomized clinical trial, or attempt to draw inferences from the data at hand. The interplay between standard evaluative criteria for screening, evidence-based medicine, and the existing evidence has been at the heart of the debate over the efficacy and value of mammography screening for women ages 40–49.

Prior to 1995, no individual trial or meta-analysis had demonstrated a statistically significant reduction in breast cancer deaths among women ages 40–49 who received an invitation to mammography screening (*7*). While a number of U.S. organizations at that time recommended that women ages 40–49 undergo mammography every one to two years, this recommendation was made on the basis of indirect evidence that mammography is beneficial to this age group (*3–4*). Other organizations did not endorse screening women ages 40–49, primarily because none of the trials up until then showed a statistically significant reduction in deaths among the 40–49 group (*5,6*). In 1995, however, Smart and colleagues published meta-analysis results that showed a statistically significant 24% reduction in deaths when all population-based trials of mammography screening were combined (*7*). More recent results reveal statistically significant mortality reductions for all trials combined, and for two individual Swedish trials (*8–10*). Thus, at this time, breast cancer screening for women ages 40–49 has met standard norms of evidence, and screening for women in their forties is endorsed by both the American Cancer Society and the National Cancer Institute. It is nonetheless important to carefully evaluate the criteria listed above and to compare the performance of screening among women in their forties and fifties according to these criteria. The remainder of this article is devoted to such an evaluation.

Disease Burden

In order to justify screening large numbers of healthy people, the disease should represent a significant public health burden. This burden may be a function of any one or combination of three disease burden measures: morbidity, mortality, and/or premature mortality. For most, breast cancer meets these criteria of

**Affiliation of author:* Cancer Control Department, American Cancer Society, Atlanta, GA.

Correspondence to: Robert A. Smith, Ph.D., Cancer Control Department, American Cancer Society, 1599 Clifton Road, NE, Atlanta, GA 30329.

importance well enough. Breast cancer is the most common malignancy diagnosed among women, and the second leading cause of mortality from cancer. The American Cancer Society estimates that in 1997, 180,200 women will be diagnosed with invasive breast cancer, 36,400 women will be diagnosed with ductal carcinoma *in situ* (DCIS), and 43,900 women will die from this disease (*11*). Breast cancer is also a leading cause of premature mortality among women, and the leading cause of premature mortality from cancer (*12*). On average, a woman who has died of breast cancer has lost 19.4 years of life she might have otherwise had (*13*). In fact, the decision to include women aged 40 and older in the Health Insurance Plan of Greater New York randomized trial of breast cancer screening was based on the observation that women diagnosed with breast cancer between ages 40–49 contributed 34% of the total years of potential life lost due to breast cancer (*14*). The emphasis on incidence rather than deaths for women in their forties is important here, since a significant proportion of the deaths that occur among women diagnosed in their forties will occur after age 50. On the basis of these early study design decisions, subsequent studies and screening guidelines by some organizations have also set the age of 40 as the earliest age at which screening should begin.

The incidence of breast cancer increases with age. The diagnosis of breast cancer is uncommon before age 25 years, and begins to increase measurably thereafter. Between ages 40–49, an estimated 1 in 66 women (1.52%) will be diagnosed with breast cancer during that 10-year period; annual age-specific incidence rates are 122.6 per 100,000 women ages 40–44 and 199.5 per 100,000 for women ages 45–49. Between the ages of 50–59 an estimated 1 in 40 women (2.48%) will be diagnosed with breast cancer in that 10-year period; annual age-specific rates are 237.1 per 100,000 women ages 50–54 and 280.0 per 100,000 women ages 55–59 (*15*). Due to trends in aging (in particular, the maturation of the postwar birth cohort), in 1997, nearly the same number of cases of breast cancer are expected to be diagnosed among women aged 40–49 as among women aged 50–59 (32,600 vs. 33,000), even through breast cancer rates among younger women are lower (*16*). In recent years, the estimated number of women diagnosed in their forties actually exceeded the estimated number of new cases among women aged 50–59 (*17*).

These measures of disease burden, taken individually or comparatively, allow one to reasonably conclude that breast cancer is an important health problem for women in their forties. While incidence is lower among women ages 40–49 compared with women in their fifties, incidence and associated measures of disease burden in each age group are sufficiently high to justify disease control efforts.

Earlier versus Later Treatment

Beyond disease burden, the disease must also meet certain criteria related to its preclinical phase (*1*). First, the preclinical condition should reasonably predict the probability of progression to clinical symptoms if left untreated. It should be noted that the preclinical condition may be invasive disease, or some important disease precursor. Diagnosis of invasive disease before the onset of symptoms is the goal of breast cancer screening. However, controversy has arisen over the increasing rate of diagnosis of DCIS resulting from greater participation in mammography, especially in younger women, on the basis that not all DCIS will progress to invasive disease. Clearly, some does, but knowledge is limited as to the proportion that will evolve into an invasive tumor. DCIS is believed to be a precursor to invasive disease for several reasons. First, it is often found in the adjacent margins of excised tumors. Second, invasive breast cancer has been shown to develop in a proportion of untreated cases (having biopsy only) of previously diagnosed benign disease, subsequently determined to be low-grade DCIS. In one study, breast cancer developed in 9 of 28 patients—five of the nine patients died of the disease (*18*). In some of the cases that did not eventually develop breast cancer, the entire lesion may have been removed at the time of biopsy, and thus effectively treated. Third, incomplete excision of DCIS has been associated with a greater probability of subsequent recurrence of invasive disease in the same area of the breast (*19*). Nevertheless, the fact that not all, and perhaps a significant proportion of DCIS may not progress to invasive disease has led to concerns regarding overtreatment, highlighted recently in an article by Ernster and colleagues (*20*). In fact, a growing clinical appreciation for the heterogeneity of DCIS has led to a number of efforts to determine prognostic factors associated with DCIS, as well as the range of treatments, some less and some more aggressive, that are appropriate based on the histologic characteristics of the disease (*21–23*). Given the current state of knowledge, reducing overtreatment of non-invasive and minimally invasive disease is a high priority. However, a diagnosis of DCIS should not be considered a "cost" of a screening program, insofar as DCIS represents a non-invasive condition with the highest probability of progression to invasive disease and thus, today, requires treatment. It should also selectively not be considered a cost only for women ages 40–49, since women ages 50–59 show a similar proportion of tumors diagnosed as DCIS (*24–25*). For individuals or the population at risk, we do not have the knowledge to tailor screening schedules in order to only detect lesions of "known" significance. Thus, it is important, however, to consider the relative importance of a diagnosis of DCIS in a screening program apart from the issue of over-treatment, especially since the latter can be addressed through professional education.

A second criterion is that the disease should have a detectable, preclinical phase, estimated as the mean sojourn time (*1,26*). The sojourn time is the estimated maximum duration of the detectable preclinical phase, and is the basis for establishing screening intervals within which beneficial lead times are attainable (*26*). The sojourn time must be of sufficient length to assure a reasonable level of disease prevalence, both for the disease to be detectable and to offer the opportunity for detection at a point when medieal intervention can make a difference in its natural history. Thus, it is axiomatic that screening intervals be less than the estimated mean sojourn time.

It has been estimated that the mean breast cancer sojourn time for women aged 40–49 is 1.7 years, whereas for women aged 50–69 it is between 3.3 and 3.8 years (*27*). This difference in estimated sojourn times has caused concern that the majority of existing trials screened women ages 40–49 at an interval that was too wide to provide the full potential benefit of an early detection program (*24,27*). Thus, the absence of a larger reduc-

tion in deaths, and the longer period of follow-up required to observe a benefit in individual trials and meta-analyses, has been attributed in large part to the failure of two-year screening intervals to adequately reduce the rate of advanced disease in women aged 40–49 (*28*).

Finally, there should be sufficient evidence that treatment for early stage disease offers significant benefits compared with treatment at a later stage. The benefits of breast cancer treatment at earlier versus later stages are well established, although on the basis of observed mortality reductions in the trials, evidence has historically been stronger for women aged 50 and older than for women aged 40–49 (*29–31*). However, since diagnosis at more favorable stages has been the basis for the observed mortality reductions in the trials, and analyses have shown similar long-term survival for women ages 40–49 compared with women ages 50 years and older when grouped by similar prognostic factors, benefits have been inferred for breast cancer detected by mammography in younger women (*31–32*). Further, longer term follow-up of trial data has revealed incremental benefits from screening among women randomized in their forties, eventually revealing statistically significant reductions in deaths after an average 12-year follow-up (*8*).

Characteristics of the Screening Test

Provided that the disease in question meets the characteristics described above, the test must meet acceptable criteria for accuracy and reliability. In other words, it must do a reasonably good job to correctly distinguish those who probably have the disease from those who probably do not have the disease. The conventional performance measures are the cancer detection rate, sensitivity, specificity, and positive predictive value. These measures are defined by end results in the context of a breast cancer screening program. By convention, the basic measurements for calculating these outcome measures are as follows: A true positive (TP) can be defined as breast cancer diagnosed within one year after a biopsy recommendation following an abnormal mammogram. A true negative (TN) can be defined as no evidence of breast cancer within one year of a normal mammogram. A false negative (FN) can be defined as a cancer diagnosed within one year of a normal mammogram. Finally, a false positive (FP) can be defined several ways, each relevant to the focus of evaluation in a screening program, and each according to the criterion that there is no evidence of breast cancer within one year after the definition of a positive finding. First, the false positive rate can be based on cases recalled for additional imaging evaluation after an abnormal screening mammogram. An alternative measure is based on the number of cases referred to biopsy or surgical consultation after an abnormal mammogram. A third definition considers only those who have actually undergone biopsy after an abnormal mammogram. Each false positive measurement, in turn, represents additional progression into the diagnostic process (*33*).

Sensitivity is a measure of the probability of detecting a cancer when a cancer exists, or the proportion of patients found to have cancer within one year of screening who were identified as having an abnormality at the time of screening. Sensitivity is estimated by TP/(TP + FN). Specificity is a measure of the probability of correctly identifying an individual as not having cancer when no cancer exists, or the proportion of patients not found to have cancer within one year of a normal screening examination. Specificity is estimated as TN/(TN + FP). The positive predictive value (PPV) varies according to the definition of a false positive result, and is the proportion of cases correctly identified as having cancer among all cases identified as positive according to the three definitions listed above (*33*). In other words, positive predictive value is given by TP/(TP + FP).

The goal of a screening program is to achieve uniformly high sensitivity and specificity, and the relative importance of accuracy for either of these measures is a function of the consequences and severity of an error, both for the individual and the cost of the screening program. From a measurement standpoint, the sensitivity and specificity of mammography are influenced by number of factors, including the quality control of the screening tests, interpretation thresholds, and the screening interval. Thus, any assessment of existing estimates must consider the characteristics of the screening program from which they derive (*34*). For this reason, constant monitoring of the performance of a screening program is essential to determine those dimensions of sensitivity and specificity inherent in the interplay between the disease and the technology at hand, and those which may be influenced by improvements in technique and operation.

How well does screening women ages 40–49 measure against screening women ages 50–59 according to these criteria? The Agency for Health Care Policy and Research's (AHCPR) Clinical Practice Guidelines No. 13: Quality Determinants of Mammography included performance measures to help mammography facilities evaluate medical audit data (*33*). According to the AHCPR guideline, if measurable, sensitivity should exceed 85%, specificity should exceed 90%, positive predictive value based on abnormal screening exam should be between 5–10%, and positive predictive value when biopsy is recommended should be between 25–40%. Data from established screening programs in the United States typically reveal that the efficiency of screening improves somewhat with age; this is especially true for positive predictive value measures, since they depend on the underlying prevalence of disease (*35*). However, in these series, and those data reported elsewhere, screening performance for women ages 40–49 and 50–59 approximates these performance measures and was more similar than dissimilar (*35–39*). Further, in a University of California, San Francisco (UCSF) program, a substantial decline in sensitivity was observed as the screening interval increased among participants in the program (*36*), meaning many of the existing measures of sensitivity from trials and other studies must be interpreted in the context of not only accuracy of interpretation, but the width of the screening interval. This is especially true when comparing sensitivity data for women aged 40–49 with older women, since women aged 50 and older are estimated to have a much wider mean sojourn time, one that is more coincident with the average screening intervals in the trials (*24,27–28*).

Moreover, data from UCSF and Albuquerque presented at the 1997 National Institutes of Health Consensus Development Conference on Breast Cancer Screening for Women Ages 40–49 showed similar performance for women ages 40–49 and 50–59 with respect to tumor size, nodal involvement, and the rate of advanced cancers (*36–37*). Other published reports have shown similar comparative performance (*38–39*). Still, Tabar and col-

leagues have argued that these conventional measures lack the necessary precision to fully assess the performance of a breast cancer screening program, and the argument is compelling in light of the varying end results and measures of sensitivity observed in previous studies (*3,40–41*). Mammographic sensitivity is not simply a measurement of test accuracy. Underlying the measurement of sensitivity is disease prevalence, characteristics of the population being screened, image quality, interpretative skill, screening intervals, and the threshold for intervention. Further, since breast cancer is a heterogeneous disease, similar measures of sensitivity are no assurance of detecting the same mix of cancers at favorable prognostic stages. For these reasons, Tabar *et al.* recommend the following benchmarks for the evaluation of the performance of a screening program: (1) more than 50% of screen-detected cancers should be smaller than 15 mm; (2) 30% or more of grade 3 cancers detected on screening should be less than 15 mm; and (3) more than 70% of cancers detected on screening should be node negative (*40*). In addition, high rates of participation are required, and participants should adhere as closely as possible to a recommended interval. More than anything else, the goal of a breast cancer screening program is a significant reduction in the rate of advanced disease over what would be expected in the absence of screening.

Conclusion

As noted above, the decision to screen is based on factors related to the importance of the disease as a public health problem and the ability of a screening test and program to meet acceptable levels of performance. Population-based screening is generally thought to be justified if the disease is important, and the screening test is judged to meet accepted criteria related to accuracy, efficacy, and practicality. While these criteria are commonly applied as an evaluative template, there are no specific thresholds by which decisions to offer or not offer screening can be made. A screening test may fail to meet any one of these criteria and therefore deemed not useful, i.e., it may have low sensitivity, or lower sensitivity than an alternative test. However, it is also the case that these criteria may be evaluated collectively, since the benefits, costs, and consequences of these criteria considered together may vary in important ways according to the population, disease, and test under scrutiny. Still, on balance, the same data may lead to different conclusions about the value of screening in a population, and decisions to recommend or not recommend screening may be more complicated when the underlying evidence is more inferential than direct. However, once the decision to screen has been reached, it is critical that screening programs are carefully monitored and that attention is devoted to using results to improve performance. In general, a breast cancer screening program must have high levels of participation and must achieve acceptable levels of performance in terms of sensitivity and specificity. More fundamentally, for screening to be effective, the program must reduce the incidence rate of advanced breast cancer in a population.

References

(*1*) Morrison AS. Screening in Chronic Disease. New York: Oxford University Press, 1992.

(*2*) Cole P, Morrison AS. Basic issues in population screening for cancer. J Natl Cancer Inst 1980;64:1263–72.

(*3*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the international workshop for screening for breast cancer. J Natl Cancer Inst 1993;85:1644–5.

(*4*) Dodd GD. American Cancer Society guidelines on screening for breast cancer: an overview. CA 1992;42:177–80.

(*5*) Smith RA. Breast cancer screening guidelines. Women's Health Issues 1992;2:212–17.

(*6*) U.S. Preventive Services Task Force. Guide to clinical preventive services, 2nd ed. Baltimore: Williams & Wilkins, 1996.

(*7*) Smart CR, Hendrick RE, Rutledge JH, Smith RA. Benefit of mammography screening in women aged 40–49: current evidence from randomized controlled trials. Cancer 1995;75:1619–26.

(*8*) Hendrick RE, Smith RA, Routledge JH, Smart C. Benefit of mammography screening in women ages 40–49: current evidence from randomized clinical trials. Presented at the NIH Consensus Development Conference on Breast Cancer Screening for Women Ages 40–49, 1997 January 21–23; Bethesda (MD).

(*9*) Bjurstam N, Bjornel L, Duffy SW. The Gothenberg breast screening trial: results from 11 years' follow-up. Presented at the NIH Consensus Development Conference on Breast Cancer Screening for Women Ages 40–49, 1997 January 21–23; Bethesda (MD).

(*10*) Andersson I. The Malmö mammographic screening trial: update on results and a harm-benefit analysis. Presented at the NIH Consensus Development Conference on Breast Cancer Screening for Women Ages 40–49, 1997 January 21–23; Bethesda (MD).

(*11*) American Cancer Society. Facts and Figures. Atlanta: American Cancer Society, 1997.

(*12*) CDC. Premature mortality due to breast cancer—United States, 1984. Morbidity and Mortality Weekly Report, 1987;36:736–9.

(*13*) National Cancer Institute. Stat Bite: Average years of life lost from cancer. J Natl Cancer Inst 1995;87:956.

(*14*) Shapiro S, Venet W, Strax P, Venet L. Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986. Baltimore: Johns Hopkins Press, 1988.

(*15*) Ries LAG, Kosary C, Hankey BF, Harras A, Miller BA, Edwards BK. SEER Cancer Statistics Review, 1973–1993: Tables and Graphs, National Cancer Institute. Bethesda, MD, 1996.

(*16*) American Cancer Society. Breast Cancer Facts and Figures, 1997. Atlanta: American Cancer Society, 1997.

(*17*) Smith RA. Epidemiology of breast cancer. In Kopans DB, Mendelson EB, editors. Syllabus: a categorieal course in breast imaging. Chicago: Radiological Society of North America, 1995.

(*18*) Page D, Dupont W, Rogers L, Jensen R, Schuyler P. Continued local recurrence of carcinoma 15–25 years after a diagnosis of low-grade ductal carcinoma *in situ* of the breast treated only by biopsy. Cancer 1995;76: 1197–200.

(*19*) Frykberg ER, Bland KI. Management of *in situ* and minimally invasive breast carcinoma. World J Surgery 1994;18:45–57.

(*20*) Ernster VL, Barelay J, Kerlikowske K, Grady D, Henderson IC. Incidence and treatment for ductal carcinoma *in situ* of the breast. JAMA 1996;275: 913–8.

(*21*) Lagios, M. Duct carcinoma *in situ*. Surgical Clinics of North America 1990:70:853–71.

(*22*) Silverstein M, Craig P, Waisman J, Lewinsky B, Colburn W, Poller D. A prognostic index for ductal carcinoma *in situ* of the breast. Cancer 1996;77:2267–74.

(*23*) Schnitt S, Harris J, Smith B. Developing a prognostic index for ductal carcinoma *in situ* of the breast. Are we there yet? Cancer 1996;77:2189–92.

(*24*) Tabar L, Fagerberg G, Chen RH, Duffy SW, Smart CR, Gad A, Smith RA. Efficacy of breast screening by age. New results from the Swedish two-county trial. Cancer 1995;75:2412–19.

(*25*) Smart CR, Byrne C, Smith RA, Garfinkel L, Letton AH, Dodd GD, Beahrs OH. Twenty-year follow-up of the breast cancers diagnosed during the breast cancer detection demonstration project. CA 1997;47:134–49.

(*26*) Walter SD, Day NE. Estimation of the duration of the preclinical state using data. Am J Epidemiol 1983;118:865–86.

(*27*) Duffy SW, Chen HH, Tabar L, Day NE. Estimation of mean sojourn time in breast cancer screening using a Markov-chain model of both entry to and exit from the preclinical detectable phase. Statistics in Medicine 1995;14: 1531–43.

(*28*) Breast cancer screening with mammography in women aged 40–49 years. Report of the Organizing Committee and Collaborators, Falun Meeting, Falun, Sweden (1996 March 21–22). Int J Cancer 1996;68:693–9.

(*29*) Hurley SF, Kaldor JM. The benefits and risks of mammographic screening for breast cancer. Epidemiologic Reviews 1992;14:101–30.

(*30*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish ran-

domised trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.
(31) Tabar L, Duffy SW, Burhenne LW. New Swedish breast cancer detection results for women aged 40–49. Cancer 1993;72:1437–48.
(32) Ries LAG, Henson DE, Harras A. Survival from breast cancer according to tumor size and nodal status. Surgical Oncology Clinics of North America 1994;3:35–50.
(33) Bassett LW, Hendrick RE, Bassford TL, et al. Quality Determinants of Mammography. Clinical Practice Guideline No. 13. AHCPR Publication No. 95-0632. Rockville (MD): AHCPR, DHHS, PHS, 1994.
(34) Chen HH, Duffy SW. A Markov-chain method to estimate the tumor progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. The Statistician 1996;45:1–11.
(35) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;271:2444–50.
(36) Sickles EA. Screening outcomes: clinical experience with service screening using modern mammography. In: Program and Abstracts. NIH Consensus Development Conference: Breast Cancer Screening for Women Ages 40–49. National Institutes of Health, Bethesda (MD), 1997.
(37) Linver MN. Mammography outcomes in a practice setting by age: prognostic factors, sensitivity, and positive biopsy rate. In: Program and Abstracts. NIH Consensus Development Conference: Breast Cancer Screening for Women Ages 40–49. National Institutes of Health, Bethesda (MD), 1997.
(38) Curpen BN, Sickles EA, Sollitto RA, Ominsky SH, Galvin HB, Frankel SD. The comparative value of mammographic screening for women 40–49 years old versus women 50–64 years old. AJR 1995;164:1099–103.
(39) Thurfjell EL, Lindgren JAA. Breast cancer survival rates with mammographic screening: similar favorable survival rates for women younger and those older than 50 years. Radiology 1996;201:421–6.
(40) Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grotoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiologic Clinics of North America 1992;30:187–210.
(41) Kerlikowske, KM. Outcomes of modern screening mammography. In: Program and Abstracts. NIH Consensus Development Conference: Breast Screening for Women Ages 40–49. National Institutes of Health, Bethesda (MD), 1997.
(42) Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the Mammography Quality Standards Act (MQSA). AJR 1995; 165:19–25.

Study Design of Randomized Controlled Clinical Trials of Breast Cancer Screening

*Eugenio Paci, Freda E. Alexander**

Evaluation of population screening must be based on a randomized clinical trial (RCT) with the study population randomized into two arms: an intervention group invited to screening and a control group not invited to screening. Reduced mortality in the intervention group is evidence of a benefit from screening. Individual randomization is the ideal, but cluster randomization is often used for logistical and ethical reasons. The use of volunteer subjects is methodologically acceptable, but results cannot be generalized. Seven RCTs of breast cancer screening by mammography have been carried out in the United States, Canada, Sweden, and Scotland. All the studies, except the Canadian, were designed to assess the effect of screening across a wide range of ages at entry. The question of the efficacy of breast cancer screening at younger ages (< 50 years) arose early, after the first results were reported. To address this question, basic elements of the screening protocol must be considered when interpreting the results; these are screening modality (e.g., mammography with or without physical examinations), interscreening interval, and number of screening rounds. This article examines the possible influence of these factors and reviews the design choices and the characteristics of the seven RCTs. [Monogr Natl Cancer Inst 1997;22:21–25]

Randomized Controlled Trials of Screening for Cancer

Three biases arise in the evaluation of screening for early detection of cancer when screen-detected cases are compared with other cases: selection bias, lead-time bias, and length bias. The first arises when people who accept an offer of screening are compared with those who refuse such an offer, since it is impossible to know what factors lead to such a decision and how these factors affect other health behaviors that may, in turn, influence the chance of dying from the cancer in question. The other two biases arise when screen-detected cancers are compared with other cancers. If survival from time of diagnosis is taken as an endpoint, then, since screening has advanced the diagnosis, the survival time will appear to be longer even if the time of death has not been changed—this is lead-time bias. In addition, the total series of screen-detected cancers will, on average, develop more slowly during the preclinical phase (so that they spend a longer time in that phase), and they might also be those which would continue to progress more slowly following clinical symptoms—this is length bias.

The only valid method of avoiding these biases in the evaluation of population screening (e.g., for breast cancer) is the use of randomized controlled trials (RCTs) (*1*). The basics of the design of these trials are well established. A study population is identified and randomized into two arms—one that receives an invitation to screening (the intervention arm) and one that does not receive an invitation (the control arm) under the protocol to be evaluated. All other health care and therapy should be independent of the study arm. The entire study population is followed up for a (usually lengthy) time period, after which disease-specific mortality in the two arms of the trial from the date of randomization to the end of follow-up is compared. Reduced mortality in the intervention arm is evidence of the beneficial effect of screening.

A number of basic design features deserve attention and are discussed below in the context of breast cancer screening.

The Identification of the Study Population

Women who have already been diagnosed with breast cancer cannot benefit from screening and so are invariably excluded from the study population, although identification of these ineligible women is not always straightforward. Once ineligible women have been excluded, the study population is usually 1) a geographical population or one which is representative of this, or 2) subjects who have volunteered to participate in an RCT. Volunteers, in turn, can be solicited in two ways: In the first case, consent is given by the intervention group after randomization (''single consent'' design), and usually only by those to decide to attend screening; in the second, consent is given prior to randomization, in which case both the intervention and eontrol groups will have given informed consent once these groups are randomized (''double consent'' design). Double consent has been rare in trials of breast cancer screening but is now frequently used for trials of other cancer screening (*2*). The main advantage is that rates of acceptance of the offer of screening in the intervention arm will be higher. Disadvantages (see below) include increased possibility of contamination (i.e., use of other screening facilities) in the control arm, difficulties in ensuring that randomization is blind, and a possible perceived professional obligation to provide some minimal screening for all who have volunteered—that is, to the entire study population.

A further problem with volunteers is that the results may not be generalizable to any other geographical population, since those who volunteer in one locale may differ, for example, in their health awareness and breast cancer risk compared with

**Affiliations of authors:* E. Paci, Epidemiology Unit, Center for the Study and Prevention of Cancer, Azienda Careggi, Florence, Italy; F. E. Alexander, Department of Public Health Services, University of Edinburgh, Edinburgh, U.K.

Correspondence to: Dr. Eugenio Paci, Epidemiology Unit, Center for Study and Cancer Prevention, Via di San Salvi 12, 50135 Florence, Italy.

those in another region. With both choices, temporal or regional changes in the underlying breast cancer incidence, stage at presentation, or survival rates may determine that results of trials cannot necessarily be generalized.

Randomization and Blinding

In RCTs, individual randomization is the ideal, but logistical and ethical issues arise when large populations of healthy individuals are involved. For example, women invited to screening will wish to discuss with their general practitioner (GP) whether or not to accept; if only half of a particular GP's patients have been (randomly) invited, this may cause practical problems to the GP and lead to resentment on the part of women who were not invited. Many trials of breast cancer screening have therefore used cluster randomization by place of residence or medical practice (*3*). This, however, can reduce the efficiency of the randomization, since the number of units randomized may be drastically reduced. In the Edinburgh trial, for instance, over 40,000 women were randomized, but the randomization process was based on just 78 clusters, and biases between the two arms of the trial have been noted (*4*).

Another basic requirement of RCTs of therapy is that randomization should be blind; that is, the allocation to one arm of the trial should be conducted by someone without knowledge of clinical characteristics of the arm which may influence prognosis. The same criteria must apply in trials of screening: that is, those conducting the randomization must have no knowledge of characteristics that might influence a subject's chances of dying of the cancer (and, hence, of being diagnosed with it). Such characteristics might include breast cancer risk factors, physical symptoms, socio-economic status, and so on. For trials with a "geographical" study population (as defined above), this presents no problem, but when the study population consists of volunteer subjects who may have had some medical examination prior to consent and prior to randomization, it is essential that blindness is achieved.

Contamination, Compliance, and Prescreening

The benefit of screening can only apply to women who are screened when compared to those who are not. Maximum effect would be seen if all women in the intervention arm and none in the control arm were screened. This would, in fact, provide an accurate estimate of the benefit of screening. However, the observed differences between the two arms of the trial will give diluted effect estimates (accompanied by loss of statistical power) if either of the following occur: women in the intervention arm do not accept the offer of screening (low compliance) or women in the control arm find alternative sources of screening (contamination). Quantifying compliance is relatively straightforward although estimating its impact is difficult; quantifying contamination is almost impossible since, even if one counts the numbers of the appropriate tests done on members of the control population, these tests may have been done on account of symptoms (in which event they are part of usual health care and do not cause contamination). However, if an initial medical examination is given to all members of the study population and includes an element of screening for the cancer (e.g., a clinical examination of the breasts), the effect will be similar to that of contamination.

Finally, women in the intervention arm (especially screen-detected cases) may be more likely to be treated in specialist centers, and this has the potential to introduce confounding of trial arm by treatment and by other factors that influence the survival experience (*5*) independently of screening. This may be unavoidable, but monitoring is mandatory.

Statistical Power and Subgroup Analysis

The statistical power of RCTs of screening, as of therapy, is based on the number of events expected in the two arms of the trial. Since, unlike therapeutic trials, the study population is initially disease free, this leads to a requirement for both very large numbers (25,000–100,000 or more) in the study population and long-term (seven years or more) follow-up (*6*). These numbers are needed to provide adequate statistical power to detect an effect; higher numbers are required to provide precise estimates of the effect (i.e., narrow confidence intervals) and to permit adequate power for subgroup analyses (e.g., women below age 50 at entry).

Endpoints

The endpoint of interest in screening RCTs is, invariably, disease-specific death (or an estimate of this derived from the use of 'surrogate' or 'interim' endpoints [*7*]). The ascertainment of all relevant deaths and validation of their status are critical in the design of screening RCTs. It is possible for biases to arise at both points and essential that this be avoided. Those women who have attended for screening will be followed up at the times of future screening visits and, if cancer is detected, may be treated in an associated unit, so that ascertainment of subsequent death, if it occurs, is straightforward. The only likely method of ascertaining deaths in the control arm and among women who do not attend when invited to screening is record linkage or 'flagging' with national death registers. Information from such sources must be complete if bias is to be avoided. The cause of death must be taken either from an entirely objective source (e.g., death certification) or must use appropriate blinding of those reviewing. There is now good evidence that use of death certificate information does not lead to error in statistical analyses, although individual errors may occur (*8*).

Objectives

Finally, we note that there is a tension between two objectives of screening trials. The first (as for Phase 2 clinical trials) is to determine whether screening can reduce mortality; this requires optimal performance of maximal screening in terms of frequency (number of years between routine invitations to screening), screening methodology (e.g., number of mammographic views, qualifications of readers, and use of duplicate reading), biopsy decisions (i.e., protocol used to select for biopsy), and so forth. The second is to provide information that can be interpreted in terms of disease natural history and cost-effectiveness. These two do not always lead to the same design choices.

RCTs of Breast Cancer Screening

Seven RCTs have been carried out in the United States, Sweden, Canada, and Scotland to assess the efficacy of breast cancer

screening (Table 1). The first RCT, the Health Insurance Plan (HIP) study, was launched in December 1963 in order to determine "whether periodic breast cancer screening with mammography and clinical examination of the breast holds substantial promise for lowering mortality in the female population from breast cancer" (*9*). The HIP study was designed to assess the effect of screening independently of the age at entry; nevertheless, the possibility of a lower efficacy of screening by age was immediately evident, although the interpretation of data was difficult because of small numbers. During the early seventies—the period of the planning phase and start of the Swedish trials—it became increasingly clear that the impact of screening was different in younger women. This observation has influenced the design of all trials since the HIP study. In the Malmö trial (start: 1976) the age at entry was postponed to 45; in the Two County Study (TCS) the age range at entry was 40–74, but the interscreening interval was shorter (24 months) for women 40–49 years old at entry. However, the Canadian trial (NBSS-I), which began in 1980, was the only study specifically designed to examine whether screening of younger women was effective.

Study Population Identification

All but one RCT used a geographical (or representative) population with the single-consent design and no examination of women in the control arm. Only NBSS-I used a double-consent design with a volunteer study population. All women enrolled in both arms of that trial were given a physical examination before randomization; the results of this examination did not influence eligibility. The HIP study is considered comparable to a population-based study, although the population at issue was not defined on the basis of a demographic population list. In Sweden, the population list was based on the Municipality Registry and, in Edinburgh, on the General Practitioner patient lists. Both of these lists cover the total resident population.

Randomization

The randomization procedures varied from trial to trial. The TCS adopted a cluster randomization based on geographical and administrative areas. The Edinburgh trial also adopted cluster randomization with the random unit being the general practice. Other trials were randomized individually (or used a systematic procedure approximating this, as in Gothenburg and Stockholm). The issue of blindness was critical in NBSS-I because randomization came after the clinical examination of the volunteer population. Breast cancer cases detected at the initial physical examination were included (by design) in the published analysis.

Screening Protocol

Three basic elements are especially relevant in the comparison of the breast cancer screening protocol between trials:

1) the screening modality (the number of views used for mammography and whether a clinical examination was included);
2) the interscreening interval (the time between routine screenings for the intervention group); and
3) the number of rounds (the number of occasions on which routine screening was offered to the intervention group).

Table 2 shows the main characteristics of the screening protocol adopted in each trial.

Mammography was carried out in two standard views in the HIP study and in NBSS-I. In the largest trial carried out in Sweden, the TCS, only one oblique, single-view mammography was performed. In the Malmö and Edinburgh trials, and in the most recent Swedish trial, the Gothenburg trial, two-view mammography was scheduled at the prevalence (first) screening, and an oblique, single-view mammography was used at subsequent rounds.

The initial physical examination of the breasts was included by design for all women recruited into the NBSS-I trial. All trials conducted outside Sweden included regular clinical examination for women in the intervention group. The protocol of the Edinburgh trial included four biennial mammography examinations and annual clinical examinations over the same period. The Swedish trials were based on mammography only.

Whether physical examination should be used in addition to mammographic screening has been debated for a long time, with differing opinions in America and Europe (*10*). Generally, European screening guidelines for older women (aged 50 or more) include only mammography. In the Edinburgh trial, it was estimated that the proportion of cases detected by screening would have been reduced by 5% if the physical examination had been omitted (*11*).

The HIP study planned four screening rounds for the women in the invited group, and women were actively followed up after the end of the screening schedule. The number of rounds for the invited groups varied in the other trials from four to five (except the Stockholm trial, which stopped after two). After that, women

Table 1. Characteristics of breast cancer screening randomized trials (ages 40–49)

Start	Study name	Population	Age range	Invited group	Control group	Randomization@
1963	HIP	Pop.*	40/49	14,432	14,701	I
1976	Malmö	Pop.	45/49	3,795	3,769	I
1977	TCS	Pop.	40/49	19,844	15,604	C
1979	Edinburgh	Pop.	45/49	11,370	10,269	C
1980	NBSS-I	Vol.	40/49	25,214	25,216	ø
1981	Stockholm	Pop.	40/49	14,842	7,103	I+
1982	Gothenburg	Pop.	40/45	10,821	13,101	I+

*Nondemographic population.
@Randomization prior to consent except where noted: I = individual; C = cluster; ø = physical examination prior to randomization; + = systematic procedure equivalent to individual randomization.

Table 2. Screening protocol of the breast cancer screening randomized trials (ages 40–49)

Start	Study name	Number of views*	Physical examination	Interscreening interval	Number of rounds
1963	HIP	2	yes	12	4
1976	Malmö	2,1	no	21	8
1977	TCS	1	no	24	4
1979	Edinburgh	2,1	yes	24@	4
1980	NBSS-I	2	yes	12	5
1981	Stockholm	1	no	28	2
1982	Gothenburg	2,1	no	18	5

*2,1 = two views at the first, one view at subsequent screening.
@12-month interval for physical examination.

in both arms were invited to have mammography as service screening. Follow-up is still ongoing in all these trials.

The interscreening interval varied between trials. The HIP and NBSS-I trials had a one-year interval. The interval was longer in the Swedish trials, ranging from 18 months in the Gothenburg trial, to 21 months in the Malmö trial, to 28 months in the Stockholm trial. In the Gothenburg trial, the last trial started in Sweden, the interval was 18 months.

Indicators like sensitivity, specificity and program predictive value estimated from the occurrence of screen-detected, interval, and clinically detected cancers in the trials or in observational studies have been used to compare the performance of different breast cancer screening programs (*12*). Based on TCS data and using statistical models, Tabar et al. (*13*) have estimated the relationship between surrogates and observed mortality reduction. Their conclusion is that the interscreening interval should be shorter in younger women (ideally, one year), since the lead time is shorter.

Characteristics of Breast Cancer Cases

All the trials, although designed to address mortality reduction, have collected information on the characteristics of breast cancer cases. The main indicators considered relevant for the evaluation of screening process—detection rate and interval cancer rate—have been published from all the trials. The pathologic classification of cases (pTNM, grade) varied in different trials, but data were not collected according to specific protocols. In the HIP study, data on the pathologic characteristics of the cases were available retrospectively. Up to this point, only the TCS has generated information rich enough for an in-depth evaluation of the relationship between the characteristics of tumors and the mortality reduction.

Knowledge of the pathologic characteristics of cancers occurring in the invited-to-screening population offers the opportunity of an early evaluation of the screening impact and is crucial for the evaluation of breast cancer screening programs. The underlying population stage/grade distribution has a major impact on the screening efficacy—for instance, if most cases are symptomatically Stage II, then screening that advances the diagnosis from Stage III to Stage II will have little effect; however, if 30% of cases are Stage III or worse, such screening will be beneficial. The epidemiology of the ductal carcinoma *in situ* (DCIS) has also increased rapidly in recent years because of mammographic screening in older women; DCIS is also associated with a high percentage (60%) breast cancer cases in younger women (*14*). Screening programs should therefore be required to monitor DCIS occurrence to evaluate the possible overdiagnosis and overtreatment associated with screening.

Assessment

The proportion of women recalled for assessment because of a positive finding at the screening test is a fundamental parameter for the evaluation of the human and economic cost of screening. As reported by Rutqvist (*15*) at a recent conference in Falun, Sweden, the percentage of younger women recalled for assessment ranged from 4% to 6% in the Swedish trials and, among these women, 0.2% to 0.9% were referred for biopsy. The propensity to recall women and the preference for a particular assessment modality (e.g., fine needle aspiration) differ in America and Europe for professional and cultural reasons (*10*).

The possible impact of more (or less) aggressive behavior on interval cancer rates has been studied in two retrospective analyses of interval and screen-detected breast cancer cases in Nijmegen and Florence (*16,17*). In younger women, 48% of interval cases were occult at the previous screening, with minimal signs present in 22%. Findings were very similar in the two case series.

Research on the psychosocial consequences of breast cancer screening is limited and the study of the possible adverse effects of screening have been studied only occasionally.

Risk Factors

Of the RCTs carried out until now, only the HIP and NBSS-I have published data on the risk profile of the enrolled women. Selection of traditionally high-risk groups for screening has never been considered feasible. Progress in the study of family history and cancer susceptibility genes might have important consequences for predictive genetic testing and for mammographic surveillance of high-risk groups. However, selective screening can only be considered for women at exceptionally high risk, and little is known about the efficacy of screening for women at special risk of cancer (*18*). There is also concern about radiologic screening among highly susceptible groups, such as the ataxia-telangiectasia carriers (*19*).

Discussion

The RCTs carried out until now were designed to solve the question of the efficacy of screening independently of age, and only the NBSS-I specifically studied younger women.

We have provided details of all the trials that have been conducted worldwide to evaluate mammographic screening for reducing breast cancer mortality among women with breast cancer, including women under 50 years of age at entry. To address the efficacy of screening in this age group, a total study population of 100,318 invited and 89,763 control women has been assembled, and the years of follow-up vary from 10 to 18 since the start. Despite these large numbers of women, the number of breast cancer deaths—251 in the invited group—is relatively small, so that statistical power remains limited.

Altogether there have been just seven trials, and we have described their individual characteristics against the 'ideal' outlined at the beginning of this paper. The trials differ between themselves in many important, or potentially important, respects; in practice, these are not independent, so that, for example, it is impossible to examine the effect of changing one factor (e.g., interscreening interval) while holding all others fixed. It follows that meta-analyses and overview analyses of these trials encounter many problems found when conducting similar analyses of observational studies—for instance, the interpretation of protocol or design differences across studies (*20*). In her meta-analysis of the breast cancer screening RCTs, Kerlikowske assessed the influence of protocol choices on the observed mortality reduction (*21*). The summary relative risk presented in that paper did not suggest a statistically significant influence of the different study design options.

The quality of mammography presents a particular problem and has always had a great relevance in the interpretation of the

trials. Quality has changed considerably over the last 20 years and, because of technical modifications, the comparison between older and more recent trials is difficult. In addition, quality at any one point in time may differ between trials. It has been argued that results based on historical technology are not necessarily applicable to best current practice; while this is undoubtedly true to some degree, the only evidence-based conclusions about the long-term benefit of a health intervention must rely on evaluation of benefits of best historical practice. At the same time, the different performance of mammography in younger women, both premenopausal and perimenopausal, has been documented and studied. The combined impact of mammographic technical inadequacy and breast cancer characteristics in younger women has been identified as a possible contributor to the lower efficacy of screening in younger women evident in these seven RCTs. For these reasons, quality assurance studies of mammographic screening in all future trials must be considered high priority.

Two other methodological quality issues concern the number of mammographic views taken and the inclusion of a clinical examination to complement mammography in the intervention group. A recently published United Kingdom randomized trial (a second generation trial) comparing one-view versus two-view mammography has shown that the two-view test performed better, achieving higher detection rates (+24%) and lower recall rates (–15%) than the one-view test (*22*). The trial enrolled women invited to screening after the age 50, but the findings are probably generalizable to younger women, for whom the performance of mammography is considered lower.

Both the extra mammographic views and the additional clinical examination may improve screening. It is inconceivable that they should lead to screening having less impact on breast cancer mortality, although they may be detrimental in other areas (e.g., causing higher recall and biopsy rates). These additional factors are in general present in trials with smaller mortality reduction and absent in those with higher; this direction of effect cannot be explained by the inclusion of the additional factors. It follows that differences on these criteria are not a problem when combining the trial results.

There is a further problem in interpreting results of trials of screening of younger women, since the consensus is that mammographic screening given to women of 50 years and over is efficacious. This was first pointed out when HIP results were being interpreted: the long-term benefits that eventually emerged in the intervention arm for women who entered the trial under age 50 were restricted to women who were diagnosed when they were over the age of 50 years (*9*). The critical question is whether the same benefits could have been achieved for these women if their first screen had been delayed until they attained the age of 50. This question cannot be unequivocally answered by analyses of the presently available randomized trials because of their designs, although relevant data have been published for TCS (*23*), and an observational study conducted within the Edinburgh trial addresses this issue (see Alexander, current issue). New trials are essential to answer this extremely important question. Two are in progress: the UK Age Trial, which started in 1991, and Eurotrial 40, which is in its feasibility phase. Briefly, women enter these trials while typically premenopausal (ages 40–42 years); those in the intervention arm are offered annual screening with high-quality, two-view mammography during their forties; and all women in both arms of the trial will enter national service screening programs at the age of 50. Any differences between the two arms in breast cancer mortality must be attributable to the beneficial effect of screening women in their forties in addition to that available from screening beyond the fiftieth birthday.

References

(*1*) Morrison AS. Screening in chronic disease. New York: Oxford University Press, 1992.

(*2*) Schroder FH, Bangma CH. The European Randomised Study of Screening for Prostate Cancer (ERSPC). Br J Urol 1997;79(Suppl):68–71.

(*3*) Donner A, Birkett N, Buck C. Randomisation by cluster. Sample size requirements and analysis. Am J Epidemiol 1981;114:906–14.

(*4*) Alexander F, Roberts MM, Lutz W, Hepburn W. Randomisation by cluster and the problem of social class bias. J Epidemiol Community Health 1989; 43:29–36.

(*5*) Prorok PC, Byar DP, Smart CR, Baker SG, Connor RJ. Evaluation of screening for prostate, lung and colorectal cancers, the PLCO trial. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC, editors. Cancer Screening. Cambridge: Cambridge University Press, 1990.

(*6*) Day NE. Surrogate measures in the design of breast screening trials. In: Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC, editors. Cancer Screening. Cambridge: Cambridge University Press, 1990.

(*7*) Nystrom L, Larsson L, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, et al. Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials. A comparison between official statistics and validation by an endpoint committee. Acta Oncol 1995;34:145–52.

(*8*) Sainsbury R, Haward B, Rider L, Johnston C, Round C. Influence of clinician workload and patterns of treatment on survival from breast cancer. Lancet 1995;345:1265–70.

(*9*) Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer: The Health Insurance Plan project and its sequelae; 1963–1986. Baltimore (MD): Johns Hopkins University Press, 1988.

(*10*) Jatoi I, Baum M. American and European recommendations for screening mammography in younger women: a cultural divide? [published errata appear in BMJ 1994;308:45 and 1994;308:196 and 1994;308:648]. BMJ 1993;307:1481–3.

(*11*) Alexander FE. Estimation of sojourn trial distributions and false negative rates in screening programmes which use two modalities. Stat Med 1989; 8:743–55.

(*12*) Paci E, Duffy SW. Modelling the analysis of breast cancer screening programmes: sensitivity, lead time and predictive value in the Florence District Programme (1975–1986). Int J Epidemiol 1991;20:852–8.

(*13*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Screening for breast cancer in women aged under 50: mode of detection, incidence, fatality, and histology. J Med Screen 1995;2:94–8.

(*14*) Weiss HA, Brinton LA, Brogan D, Coates RJ, Gammon MD, Malone KE, et al. Epidemiology of in situ and invasive breast cancer in women aged under 45. Br J Cancer 1996;73:1298–305.

(*15*) Rutqvist LE. Summary of the characteristics of the Swedish trials. Proceedings of the Falun Conference (Sweden), 1996 March 21–22.

(*16*) Van Dijck JA, Verbeek AL, Hendriks JH, Holland R. The current detectability of breast cancer in a mammographic screening program. A review of the previous mammograms of interval and screen-detected cancers. Cancer 1993;72:1933–8.

(*17*) Ciatto S, Rosselli del Turco M, Zappa M. The detectability of breast cancer by screening mammography. Br J Cancer 1995;71:337–9.

(*18*) Smith R, Giusti R. The epidemiology of breast cancer. In: Bassett L, Jackson V, editors. Diagnosis of diseases of the breast. Philadelphia: WB Saunders, 1997.

(*19*) Boice JD Jr, Miller RW. Risk of breast cancer in ataxia-telangiectasia [letter]. N Engl J Med 1992;326:1357–8.

(*20*) Peto R. Why do we need systematic overviews of randomized trials? Stat Med 1987;6:233–44.

(*21*) Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.

(*22*) Wald NJ, Myrphy P, Major P, Parkes C, Townsend J, Frost C. UKCCCR multicentre randomized controlled trial of one- and two-view mammography in breast cancer screening. Br Med J 1995;311:1189–93.

(*23*) Tabar L, Duffy SW, Chen HH. Re: Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials. J Natl Cancer Inst 1996;88:52–5.

Periodic Screening for Breast Cancer: The HIP Randomized Controlled Trial

*Sam Shapiro**

This paper summarizes the findings of the first breast cancer screening trial, which was initiated in December 1963 to explore the efficacy of screening. Women aged 40–64 years were selected from enrollees in the Health Insurance Plan (HIP) of Greater New York and were randomly assigned to study and control groups. Study group women were invited for screening, an initial examination, and three annual reexaminations. Screening consisted of film mammography (cephalocaudal and lateral views of each breast) and clinical examination of breasts. Breast cancer and mortality from breast cancer were examined by treatment group (study vs. control) and by entry-age subgroup. By the end of 18 years from entry, the study group had about a 25% lower breast cancer mortality among women aged 40–49 and 50–59 at time of entry than did the control group. However, to a large extent the difference among the 40–49-year-olds occurred in the subgroup with breast cancer diagnosed after these women had passed their 50th birthday, and the utility of screening women in their forties is questionable. [Monogr Natl Cancer Inst 1997;22:27–30]

The Health Insurance Plan (HIP) project was initiated in December 1963 to determine whether periodic breast cancer screening with mammography and clinical breast examination holds substantial promise for lowering breast cancer mortality among women over time (15 to 20 years). Women 40 to 64 years of age with at least a year's membership in HIP were randomly assigned to either the study group or the control group. Initially, there were about 31,000 women in each group, a figure that was reduced by about 2%, primarily through the exclusion of women identified as having a prior breast cancer diagnosis (*1–3*).

Women entered the project between December 1963 and June 1966. The screening schedule included an initial screening examination and three reexaminations at annual intervals for those who screened negative at the initial examination. About 67% of the women appeared for their initial examination, many of whom participated in succeeding examinations. Women who disenrolled from HIP continued to receive free screening examinations. Control group women followed their usual patterns of care.

Each examination consisted of film mammography (cephalocaudal and lateral views of each breast); a clinical examination of the breast by a physician, usually a surgeon; and an interview for demographic and other background information. Mammography and clinical examinations were conducted independently. Later, findings were coordinated for reports to the women and their personal physicians.

Overlapping sources of information were used to identify women with an initial diagnosis of breast cancer and to identify cause of death in the study and control groups. These sources included HIP records, hospital claims files, death records in several states, the cancer registry for New York State, the National Death Index (1979 and later years), and mail surveys answered by women 5, 10, and 15 years after entry. Cause of death was determined by reviewing death certificates and hospital and physicians' records; reviewers were blinded to whether the women were in the study or control groups.

Selected Methodological Issues

Special attention was paid to avoid sampling biases that would limit the comparisons between the study and control groups. No differences were found in a survey of personal characteristics. Breast cancer rates after 10 years of follow-up, as well as mortality from all causes of death except breast cancer, were similar in the study and control groups (Table 1). There were differences, however, between the study group women who were screened and those who refused screening; the latter group had a much higher general mortality rate and lower breast cancer incidence rate, indicating the need to combine both groups in making comparisons with the control group.

The number of breast cancer cases detected was almost equal in the study and control groups at the end of five years from

Table 1. Mortality from all causes excluding breast cancer and breast cancer detection rates: 10-year follow-up after entry

	Intervals from entry		
Rate	10 yr	1–5 yr	6–10 yr
Deaths/10,000 person-yr			
Total Study	68.6	56.3	81.4
Screened	56.8	42.9	71.1
Refused Screening	93.0	83.7	102.7
Control	68.9	58.2	80.1
Breast Cancers/1,000 person-yr			
Total Study	2.11	2.05	2.18
Screened	2.24	2.26	2.21
Refused Screening	1.86	1.61	2.13
Control	2.09	1.95	2.22

Affiliation of author: Professor Emeritus, Department of Health Policy and Management, School of Hygiene and Public Health, Johns Hopkins University.

Correspondence to: Sam Shapiro, Johns Hopkins School of Public Health, Health Policy and Management, 624 North Broadway, Room 654, Baltimore, MD 21205-1901.

entry (i.e., about 1½ years after the last women were screened in their follow-up examinations). At five years, there were 304 breast cancers histologically confirmed in the study group and 295 in the control group; at six years, the numbers were 367 and 364 breast cancers in the two groups, respectively; and at seven years, there were 426 and 439 breast cancers in the two groups. Most of the results of the trial are based on the cases detected within five years; very similar results are found when the data include the breast cancers diagnosed in years six and seven.

Rules were established for assigning breast cancer as the cause of death. This was done because of the uncertainty in using death certificate information to classify underlying cause of death for research purposes. Two physicians determined whether breast cancer was the underlying cause; differences of opinion were resolved through consultation.

Results of Screening Trial

Table 2 gives the distribution of histologically confirmed breast cancers detected during the first five years from entry for the study group by source of diagnosis: 74% of the diagnosed women in this group had been screened at least once; more cases

Table 2. Breast cancer cases histologically confirmed: study group

	Number	Percent
Study (Total)	304	100.0
Screened	225	74.0
Detected on Screening	132	43.4
Interval	93	30.6
(<12 months)	(45)	(14.8)
(>12 months)	(48)	(15.9)
Refused	79	26.0

Note: Case detection for first five years after entry.

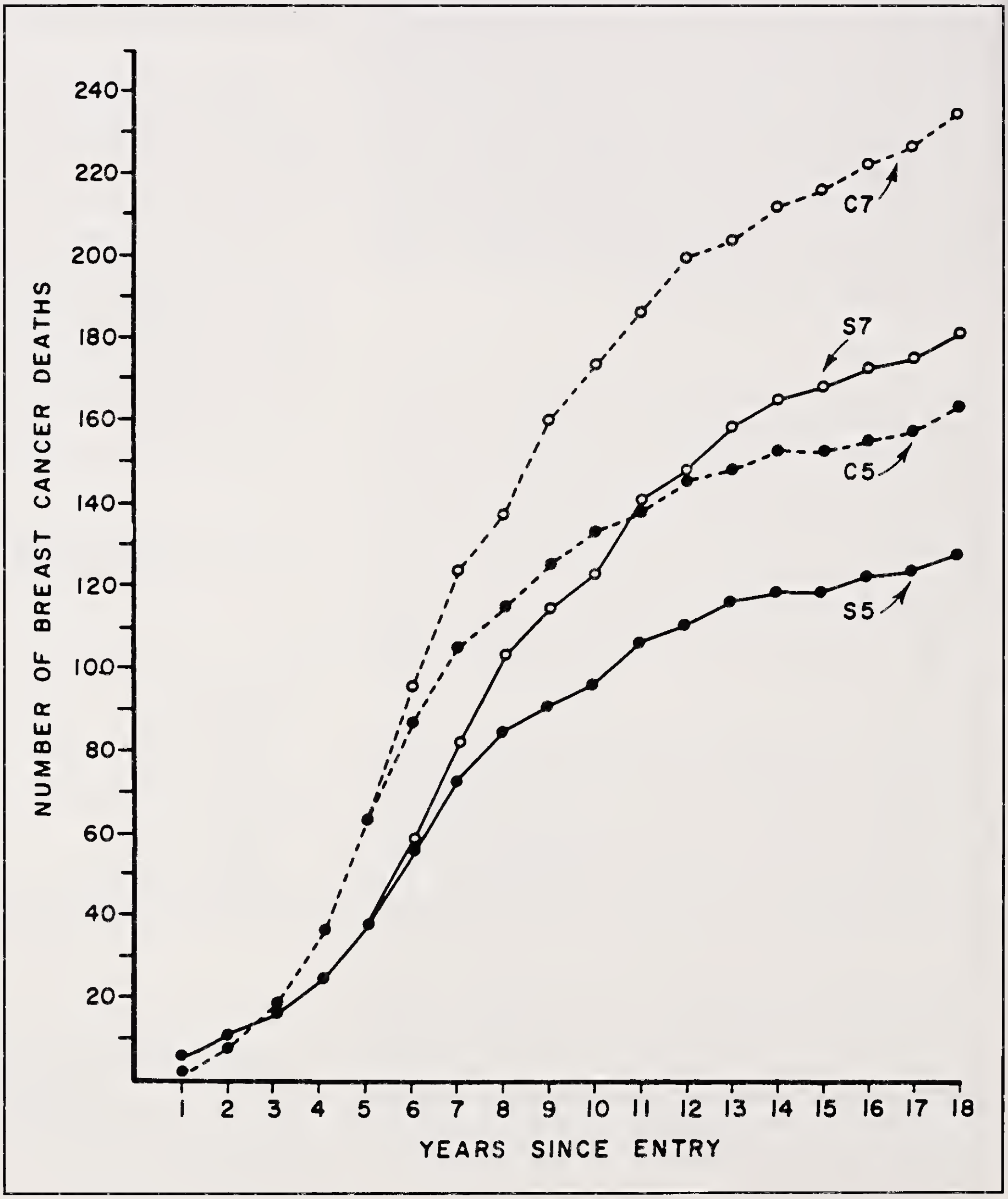

Fig. 1. Cumulative number of deaths due to breast cancer by interval since entry: all ages, study and control groups (breast cancers diagnosed within 5 and 7 years after entry).

Table 3. Percent of histologically confirmed breast cancer cases by modality of detection at screening

Modality of detection	Total	Age at entry 40–49	50–59	60–64
Total (%)	100.0	100.0	100.0	100.0
MM only	33.3	25.0	38.8	32.0
Clinical only	44.7	57.5	40.3	36.0
MM and Clinical	22.0	17.5	20.9	32.0

Table 4. Breast cancer deaths among women diagnosed in specified intervals from entry: study and control groups

Year of diagnosis after entry	Interval from entry to breast cancer death 5 yrs	10 yrs	18 yrs
1–5			
Study Group	39	95	126
Control Group	63	133	163
Percent Difference	38.1	28.6	22.7
1–7			
Study Group		123	180
Control Group		174	236
Percent Difference		29.3	23.7

were found through rescreenings than at initial examination; and about 15% of the cases were detected in the 12-month interval since the subject's last screening.

Table 3 shows the distribution of confirmed breast cancers by source of diagnosis. A higher proportion of breast cancers were detected through the clinical examination than through mammography; this was especially true for the women under 50 years of age.

As Table 4 indicates, among women aged 40–64 at entry,

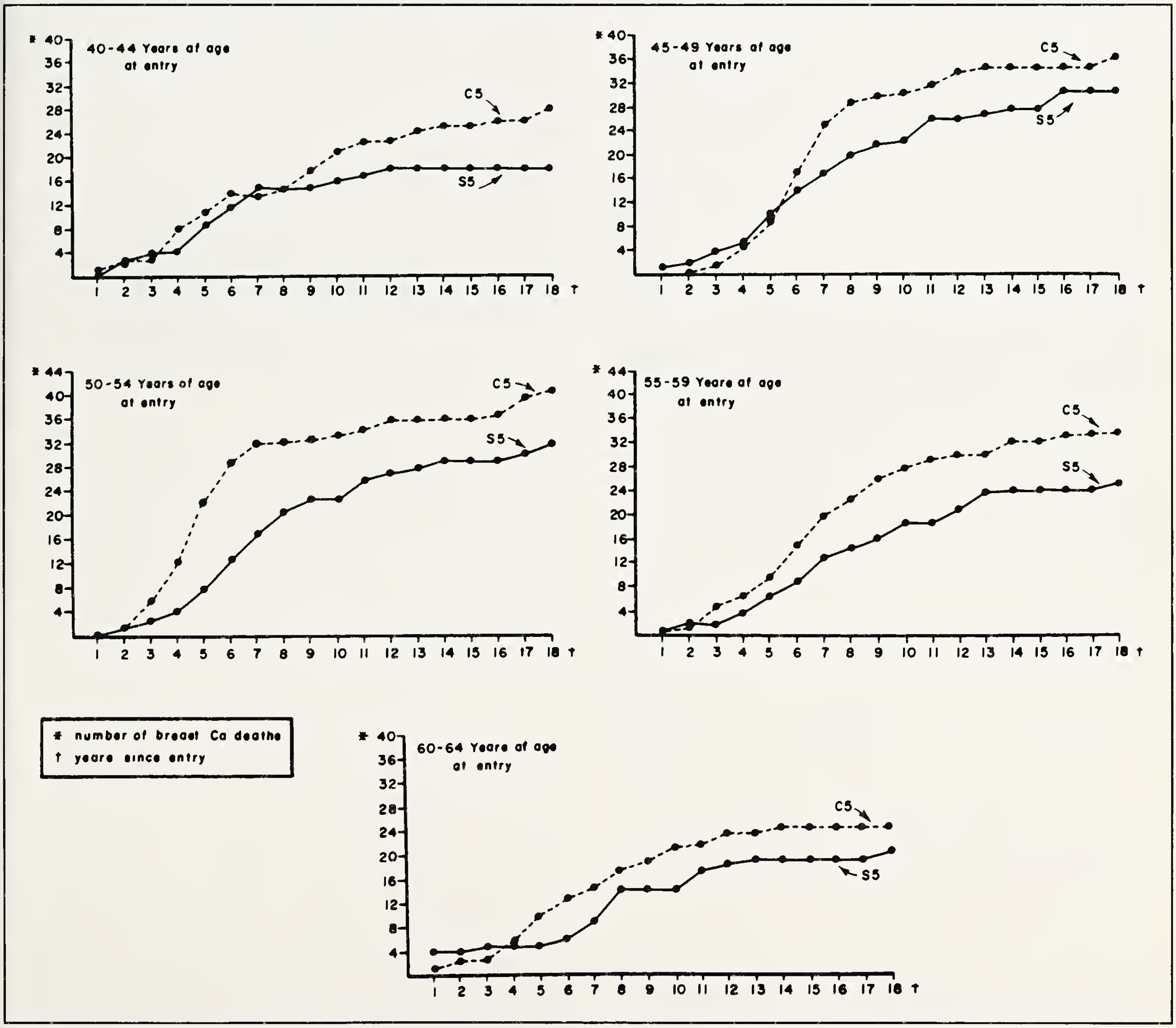

Fig. 2. Cumulative number of deaths due to breast cancer by interval since entry and age at entry: study and control groups (breast cancers diagnosed within 5 years after entry).

screening resulted in about a 30% reduction in breast cancer mortality during the first 10 years of follow-up from entry; by the end of 18 years, the reduction was close to 25%. Figure 1 plots the data for breast cancer deaths among women who had breast cancer in the first five years and in the first seven years after entry. It is clear that the same relationships apply to both sets of curves.

A favorable effect of screening appeared appreciably later among women aged 40–49 at entry than among women above this age. At 10 years from entry, mortality differentials between the study and control groups were relatively lower at ages 40–49 than at ages 50–59 but were at a similar level at 18 years of follow-up (Table 5). The delayed reduction in mortality among women aged 40–49, compared to those aged 50–59, is seen in Figure 2.

Much of the gain after 18 years of follow-up among women 40–49 is due to breast cancer cases detected when these women were 50–54. Limiting the experience to women who were still 40–49 at time of detection reduces the decrease in breast cancer mortality in this age group from 25% to 14% (Table 6). Among women 45–49 at entry and at diagnosis more deaths from breast cancer occurred in the study group (18) than the control group (14).

There are restrictions on drawing hard conclusions from these data, but the reduction in the decrease in mortality casts doubt on the ability to conclude from the HIP study that initiation of screening under the age of 50 is efficacious.

Table 5. Percent reduction in breast cancer deaths in study vs. control group women by age at entry and by selected intervals after entry

	Years from entry to death		
Age at entry	5	10	18
Total (%)	38.1	28.6	22.7
40–49	5.0	23.5	24.6
40–44	[a]	23.8	35.7
45–49	[a]	23.3	16.2
50–59	54.5	31.1	23.0
50–54	65.2	30.3	22.0
55–59	[a]	32.1	24.2
60–64	[a]	33.3	16.7

[a]Not calculated, small numbers

Table 6. Breast cancer deaths among women ages 40–49 years at entry, by age at diagnosis: study and control groups

	Deaths within 18 years from entry	
Age at diagnosis in years	Study	Control
40–49[a]	18	28
40–44	7	10
45–49	11	18
45–54[a]	31	37
45–49	18	14
50–54	13	23

[a]Age at entry: 40–44 years
[b]Age at entry: 45–49 years

References

(*1*) Shapiro S, Venet W, Strax P, et al. Ten-to-fourteen year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69:349–55.

(*2*) Shapiro S, Venet W, Strax P, et al. Selection, follow-up, and analysis in the health insurance plan study. A randomized trial with breast eancer screening, 1985. In: Selection, follow-up, and analysis in prospective studies: A Workshop. L. Garfinkel, O. Ochs, M. Mushinski, editors. NIH Publication 85-2713; National Cancer Institute Monograph 67. Washington (DC): DHHS, PHS.

(*3*) Shapiro S, Venet W, Strax P, et al. Periodic Screening for Breast Cancer: The Health Insurance Plan Project and Its Sequelae, 1963–1986. Baltimore: Johns Hopkins University Press, 1988.

The Edinburgh Randomized Trial of Breast Cancer Screening

*Freda E. Alexander**

This article presents additional follow-up analysis of women aged 45–49 from the Edinburgh Randomized Trial of Breast Cancer Screening. The screening protocol included four mammographic examinations at two-year intervals and seven annual clinical examinations. Altogether, 21,774 women aged 45–49 were recruited from 1978 to 1985 using cluster randomization. After 10–14 years of follow-up, breast cancer mortality has been reduced by 12% to 18% (rate ratios, with and without adjustment for socio-economic status, are 0.88 and 0.82 respectively, with 95% confidence intervals [CIs] of 0.55–1.41 and 0.51–1.32). These benefits are smaller than that reported previously with shorter follow-up. This article also presents data from an observational study that compared survival beyond baseline (50–52 years) of women first offered screening before and after age 50. Based on six-year data, the results suggest that earlier screening confers follow-up benefit (hazard ratio for later screening = 1.60; 95% CI: 0.96–2.67), but these findings are not statistically significant. The trial is too small to yield statistically significant results by itself, but can make useful contributions to overview and meta-analyses. [Monogr Natl Cancer Inst 1997:22:31–35]

This article presents, first off, updated data on women aged 45–49 recruited to the Edinburgh Randomized Trial of Breast Cancer Screening (ERT). The ERT initially recruited 44,288 women aged 45–64 years during the period 1978–1981. Almost all women of this age living in Edinburgh were eligible for entry to the trial. This initial sample included 11,391 women ages 45–49 years at entry (cohort 1). In addition, a further 10,383 women aged 45–49 were recruited in two cohorts during the periods 1982–1983 (cohort 2) and 1984–1985 (cohort 3). These were mostly younger women who had recently attained the age of 45 years. The average ages in these three cohorts at entry were 47.4 years, 46.1 years, and 45.8 years respectively.

The ERT methods have already been published (*1*). Important aspects of these are first, the use of cluster randomization based on primary health care units, rather than individual randomization, and second, the flagging of all women in the trial. Cluster randomization is substantially less efficient than individual randomization, since the number of units can be, as here, much smaller. Comparisons of the two arms of the ERT (offered screening versus routine health care) have revealed that the two do in fact differ, both by socio-economic status (SES) and by all-cause mortality (*5*). The intervention arm is of higher SES and has lower all-cause mortality. In addition, other specific-cause mortality is higher in the control arm for causes for which mortality rates are known to correlate positively with lower SES (*5*). As for the second key methodological feature, flagging, this allows the investigators to routinely monitor all mortality, death certificate causes of death, and the cancer registrations of all trial members, both before and after entry to the trial. Although women with a diagnosis of breast cancer before the trial entry date are not eligible for the trial, they have all been flagged.

Analyses of breast cancer mortality after seven years of follow-up (*2*) and after 10 years of follow-up (*3*) have also been published. The numbers of women and durations of follow-up were chosen to provide adequate statistical power, using a one-sided test, to detect a 30% reduction in breast cancer mortality in women offered screening. All published analyses have, however, used the two-sided tests now considered preferable. No consideration was given at the design stage to the possibility of subgroup analyses, and the trial has inadequate power to address these. Nevertheless, because of the interest surrounding the efficacy of screening women under age 50, results for women in this age group have been reported.

Women in the intervention arm were mammographically screened at entry and on three subsequent occasions for the initial cohort, two for the second, and one for the third. The fieldwork period for many of the younger (45–49 years) entrants continued into their fifties; the percentages of all mammographic screens conducted for these women while they were under age 50 were 46% (initial cohort), 79% (cohort 2) and 97.5% (cohort 3). During this fieldwork period, all women attending for mammography also received a clinical examination, which was conducted independently, and they were also invited to another clinical examination midway between two scheduled mammographic screens. Thus, the intervention protocol includes clinical examination as an adjunct to mammography. Direct quantification of the contribution of the clinical examination is not possible, but statistical modeling (*4*) has estimated that (in a steady state) use of biennial mammography alone would detect at screening 74% of all breast cancers in a screened population, and this can be increased to just 79% by the use of the clinical examination. The corresponding estimates of mean lead time with and without the clinical examination differ by just three months.

Breast cancer screening was introduced by the United Kingdom National Health Service (NHS) in 1988. This is by mammography alone, for women aged 50 years or more, and uses an interscreening interval of three years. In Scotland, this means

**Correspondence to:* Dr. Freda Alexander, Department of Public Health Science, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, UK.

See "Notes" following "References."

that women are considered for invitation every three years and receive one on the first occasion this occurs after their 50th birthday. In practice, therefore, women receive their first offers of service screening while aged 50–52 years. Women in the ERT received their first service screening offer between 1988–1990. For women in the ERT who either had been regular attendants at screening or who had been controls, these invitations were scheduled to conform to the trial design. In particular, screened women had their first service screen offered after an interval of three years from their last trial screen, and control women received their first invitation three years after they would have been invited to their last trial screen had they been in the intervention arm. All the updates were eligible to move straight into service screening after the end of their fieldwork period, and for most of these women, their first offer of service screening was at the minimum ages of 50–52 years. This necessarily dilutes the potential effect of the intervention in the updates, especially after longer follow-up periods.

The purpose of this article, then, is actually twofold. First, as noted earlier, it provides additional follow-up data on the three ERT cohorts, with and without adjustment for SES. Second, it presents data from an observational study conducted to compare women scheduled to receive their first screening (NHS or trial) before versus after age 50, to determine whether earlier screening benefits women in their fifties.

Methods

ERT Follow-Up

The women in the initial cohort of the trial completed 14 years of follow-up at the end of 1995, and sufficient time has now elapsed for death notifications to be complete. At the same time, 12 years of follow-up is complete for cohort 2 and 10 years for cohort 3. Analyses of breast cancer mortality for these longer periods of follow-up have been conducted as described previously (*3*) but restricted to one endpoint: breast cancer as the underlying cause of death according to death certificate. The use of the alternative review method in the previous report added little aggregate information, and this has been confirmed by others. The analyses, as before, adjust for the extra-binomial variation introduced by the cluster randomization using the method of Williams (*6*).

These analyses include all deaths having breast cancer as the underlying cause, whatever the time of diagnosis. In particular, the analyses include deaths of cases diagnosed after the time when NHS service screening was available to women in both arms of the trial. This is equivalent to the "follow-up" method of analysis applied to data from the Swedish two-county trial (*7*).

The Observational Study

An additional observational comparison using trial data has been conducted to compare the survival experience of women according to whether they were destined to receive their first invitation to screening while under 50 or aged 50–52 years. This can address the question, raised by the Health Insurance Plan (HIP) trial, of whether screening conducted earlier than the 50th birthday benefits women in their fifties. These analyses were applied to women who were free of breast cancer at the age of 45 years but have had breast cancer diagnosed subsequently. They are either participants in the trial or were otherwise eligible but excluded due to a diagnosis of breast cancer between their 45th birthday and their proposed trial entry date. Survival beyond a baseline age, which approximates the start of United Kingdom service screening (see below), was tested for association with the age (<50 versus ⩾50 years) when the women would have received their first offer of screening according to the trial protocol.

Three groups of women form the study population for the observational study. The first group consists of members of the intervention arm of the initial cohort of the trial; for these women, the age at entry to the trial determines whether they would (entrants 45–49 years) or would not (entrants 50–52 years) receive their first offer of screening early (<50 years). The second and third groups comprise women in the 1982–3 and 1984–5 ERT cohorts age 45–46 years at entry; for these women, the trial arm determines the age at which the first offer of screening was intended (early if in the intervention arm, ages 50–52 years if in the control arm). The comparison for the second and third groups is based on randomization. Women who would have been in one of these groups but had a diagnosis of breast cancer between the 45th birthday and trial entry date are also included.

For groups 2 and 3 in the observational study, the baseline date has been taken as the 50th birthday, and the maximum follow-up time is that used for the ERT follow-up analyses (i.e., 12 years from trial entry for the second cohort and 10 years for the third). Since these women were all considered for entry to the trial and flagged before their 50th birthday, complete follow-up information from baseline is available. To ensure complete ascertainment of deaths for women in the first group included in the observational study, it is necessary to take the 53rd birthday as baseline (by which time all had been considered for entry and flagged). For these women, the maximum follow-up period is to their 60th birthday.

Cox's proportional hazard method was used to analyze survival from baseline for women in the observational study groups 1–3 with breast cancer diagnosed between their 45th birthday and the end of the maximum period of follow-up. Deaths with breast cancer not the underlying cause have been censored, and censoring has also been imposed at two alternative endpoints: (i) six years from baseline for all women; and (ii) "variable"—that is, being the maximum available from the present data (six years for ERT cohort 3, seven years for the initial ERT cohort, and eight years for ERT cohort 2).

Both the ERT follow-up and the observational analyses (particularly the randomized comparison for the observational groups 2 and 3) are similar to that conducted in the HIP trial (*8*), where the baseline was the date of trial entry. All women in the relevant age groups and with the relevant entry times in both arms of the trial have (if free of breast cancer) had several years' opportunity of service screening, so that increased diagnosis on account of screening in the intervention arm (and prior screening in ERT cohort 1) should no longer be present. On the other hand, the ERT cannot, as did HIP (*9*), demonstrate equal cumulative incidence in the two arms because of the SES bias.

Analyses for both the ERT follow-up and the additional observational study have been restricted to flagged women, have been adjusted for cohort, and have been repeated with adjust-

Table 1. Breast cancer mortality at 14 years, women aged 45–49 at entry into ERT

	Years of entry	Years of follow-up	Breast cancer deaths	
			N	Rate/10^5 yrs.
Intervention group	1978–1981	78,761	27	3.43
	1982–1983	29,414	17	5.78
	1984–1985	31,696	2	0.63
Control group	1978–1981	75,726	33	4.36
	1982–1983	28,029	16	5.71
	1984–1985	22,662	3	1.32

Table 2. Mortality odds ratios with and without adjustment for SES for intervention arm compared with control arm. Breast cancer mortality at 14 years

Age at entry	Entry dates	Odds ratio (95% CI)	Adjusted odds ratio (95% CI)
45–49	1978–1981	0.77 (0.43–1.37)	0.84 (0.48–1.49)
45–49	1978–1985	0.82 (0.51–1.32)	0.88 (0.55–1.41)
45–49	1978–1981	0.77 (0.37–1.62)	
45–49	1978–1985[1]	0.78 (0.46–1.31)	

[1] *See* "Notes" section.

ment for SES of the primary health care unit as described in (*2,5*). The statistical packages SAS and EGRET were used to perform the analyses.

Results

The breast cancer mortality for the two arms of the ERT trial and by entry year for women aged 45–49 years at entry is shown in Table 1. The total number of deaths remains very small.

Formal comparisons of the younger entrants, when adjusted for SES, give estimated reductions of 12% to 16%, which differ little from corresponding analyses of the whole initial cohort (Table 2). None of these results are statistically significant, and all confidence intervals are wide. The point estimates of benefit are smaller than those that did not adjust for SES and those for the 10-year analysis that did not adjust for SES because of significant interaction.[1] Figure 1 shows that the difference between the intervention and control populations is absent up to six years of follow-up and then largest in the period 8–11 years. Cumulative all-cause mortality remains uniformly higher in the control group (Fig. 2).

The observational study (Table 3) shows differences in survival (from baseline) between women for whom an earlier offer of screening was and was not available. These are of borderline statistical significance. The point estimate of the hazard substantially exceeds unity in both groups and by a larger amount for groups two and three in the observational study, although the confidence intervals are very wide when subgroups are analyzed.

Discussion

The ERT mortality analysis at 10–14 years and its relation to earlier analyses are broadly in line with results from other randomized trials. Screening given during a limited fieldwork period cannot confer an unending benefit. In addition, young entrants in both the ERT follow-up and the observational study all benefited from service screening, and for groups 2 and 3 in the observational study, this generally occurred without any intervening period without screening beyond that imposed by the NHS age criteria. Thus, they all received benefit from screening

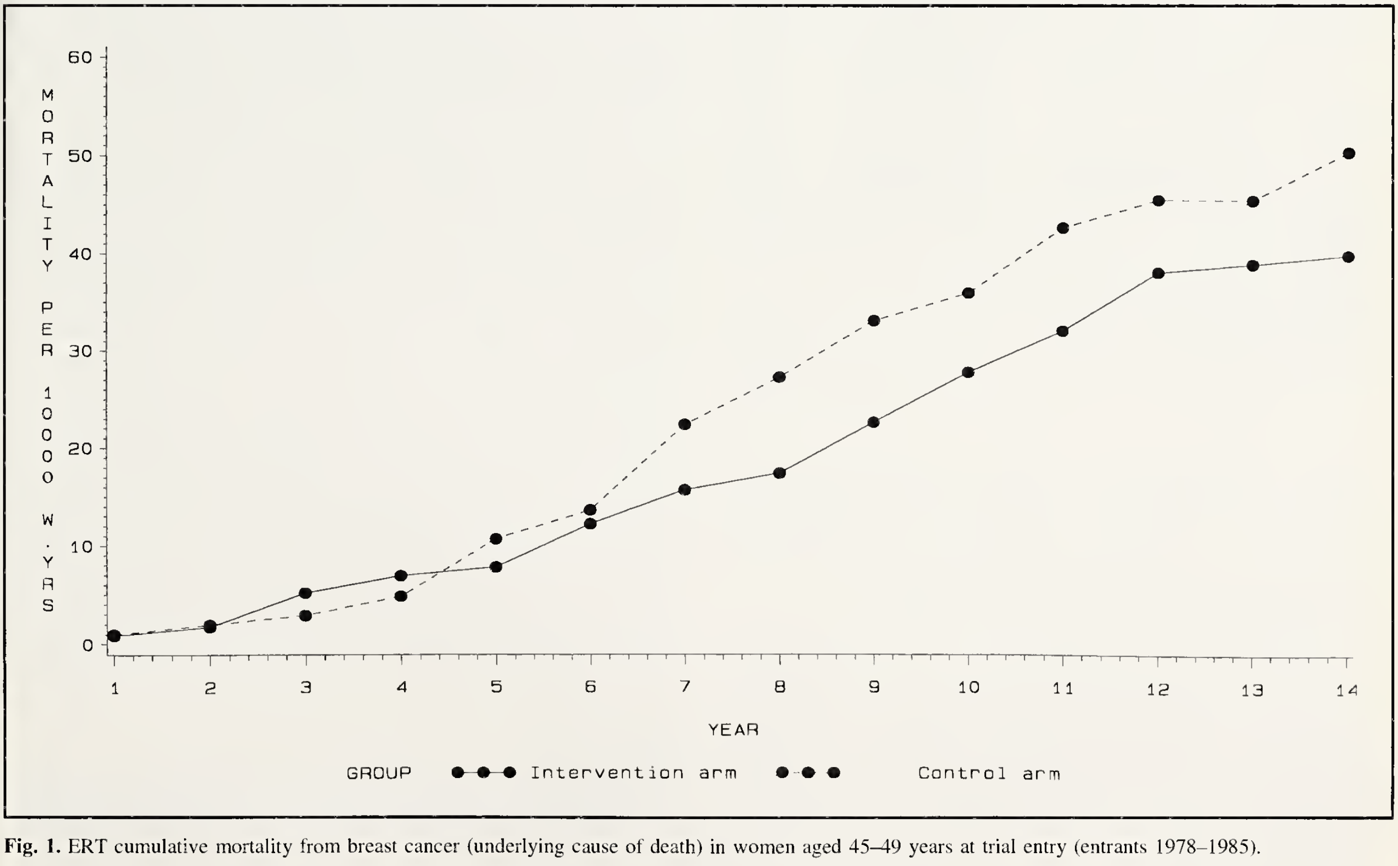

Fig. 1. ERT cumulative mortality from breast cancer (underlying cause of death) in women aged 45–49 years at trial entry (entrants 1978–1985).

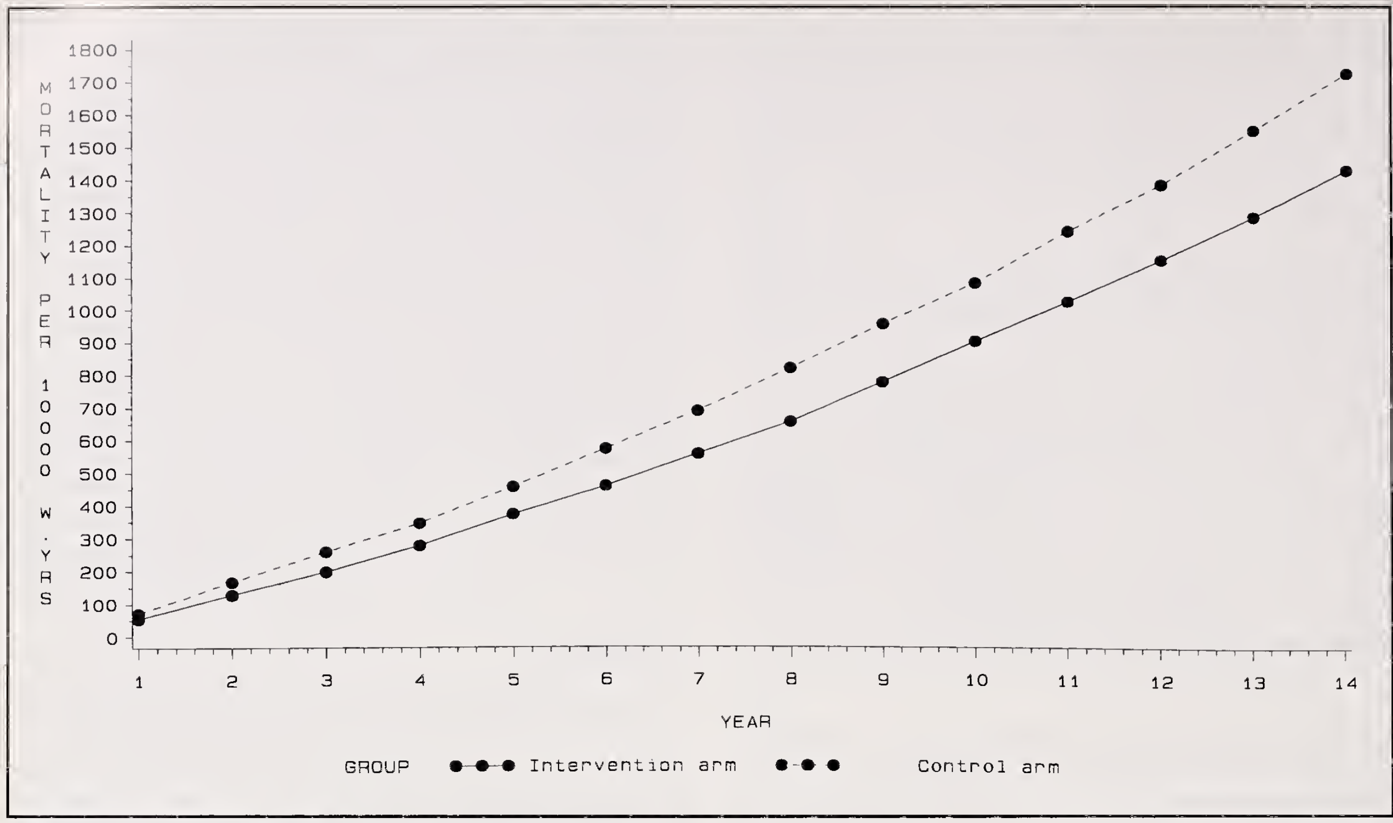

Fig. 2. Cumulative all-cause mortality in women in the initial ERT cohort.

Table 3. Results of observational study. Survival beyond baseline[1] by intended age[2] of first offer of screening

Follow-up period	Group analyzed	Hazard ratio (95% CI)[3]	*P*-value
6 yrs	all	1.60 (0.96–2.67)	0.07
	1982–1985 entrants	1.83 (0.85–3.91)	0.12
	1978–1981 entrants	1.43 (0.71–2.87)	0.31
Variable[4]	all	1.45 (0.94–2.23)	0.09
	1982–1985 entrants	1.54 (0.79–2.98)	0.21
	1978–1981 entrants	1.37 (0.78–2.42)	0.28

[1]Baseline age: age 53 years for 1978–1981 entrants. Age 50 years for other entrants.

[2]Reference group: first offer of screening <50 years.

[3]Hazard ratios are for first offer of screening to women 50–52 years with no breast cancer diagnosis up to that time.

[4]Variable follow-up: 7 years from age 53 for 1978–1981 entrants, 8 years from age 50 for 1982–1983 entrants, 6 years from age 50 for 1984–1985 entrants.

conducted at those ages recommended unequivocally by expert scientific opinion. The point estimates and their change with time and age can readily be explained this way. On the other hand, chance can explain all these patterns and the results are consistent with the null hypothesis of no benefit.

The SES bias and the difference in all-cause mortality between the two arms of the ERT follow-up trial is a cause for concern, but it has been argued previously that the effect should be conservative. Breast cancer incidence is higher in women of higher SES (*9*), but survival from time of diagnosis is longer for these women (*10*). It is *a priori* uncertain whether and in which direction breast cancer mortality will be associated with SES, although most authors assume that it will be higher in women of higher SES. This has been confirmed in the control women of the ERT (McCafferty, manuscript in preparation). The bias introduced by the cluster randomization should therefore be conservative. A special check has been made of vital status of all breast cancer cases in the 1984–1985 cohort (cohort 3); the low breast cancer mortality in this cohort is not attributable to artifacts of ascertainment of deaths from flagging.

The ERT was, as indicated above, not designed for analyses by age or other subgroups. Its importance to those considering effects of screening women under age 50 years is most likely to come from overview and meta-analyses. Results of the analyses of breast cancer mortality in the whole trial population after 14 years of follow-up and with alternative methods of SES adjustment will be published shortly (Alexander et al., manuscript in preparation).

The comparability of groups analyzed in the observational study is not justified by randomization, and confounding effects by an unknown factor cannot be ruled out. These results are preliminary, since dates of breast cancer diagnosis are still being sought for a small number of women known to have been diagnosed before trial entry (but not at present known to be before or after the 45th birthday). Although, the results do not reach conventional levels of statistical significance, they suggest that women in their fifties receive benefit from screening conducted earlier.

References

(*1*) Roberts MM, Alexander FE, Anderson TJ, et al. The Edinburgh Randomised Trial of Screening for Breast Cancer: description of method. Br J Cancer 1984;47:1–6.

(2) Roberts MM, Alexander FE, Anderson TJ, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241–6.
(3) Alexander FE, Anderson TJ, Brown HK, et al. The Edinburgh randomised trial of breast cancer screening: results after 10 years of follow-up. Br J Canc 1994;70:542–8.
(4) Alexander FE. Estimation of sojourn time distributions and false-negative rates in screening programmes which use two modalities. Statist Med 1989;8:743–55.
(5) Alexander FE, Roberts MM, Lutz W, Hepburn W. Randomisation by cluster and the problem of social class bias. J Epidemiol Comm Hlth 1989;43:29–36.
(6) Williams DA. Extra-binomial variation in logistic linear models. Appl Stats 1982;31:144–8.
(7) Nystrom L, Rutqvist LE, Wall S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993;341:973–8.
(8) Aaron JL, Prorok PC. An analysis of the mortality effect in a breast cancer screening study. Int J Epidemiol 1986;15:36–43.
(9) Sharp L, et al. Cancer Registration Statistics Scotland, 1981–1990. Information and Statistics Division, Scottish Health Services Common Services Agency, Edinburgh, 1993.
(10) Schrijers CT, et al. Deprivation and survival from breast cancer. Br J Cancer 1995;72:738–43.

Notes

[1] Alternative methods of SES adjustment are now being evaluated, and these suggest that the unadjusted results presented here may be preferable.

I would like to thank my colleagues in Edinburgh for their assistance, especially Dr. Helen Brown, Dr. Tom Anderson, Professor Sir Patrick Forrest, Dr. Alastair Kirkpatrick, Mrs. Alice Smith and the late Dr. Maureen Roberts. I also acknowledge support for the ERT from the Scottish Home and Health Department.

The Canadian National Breast Screening Study: Update on Breast Cancer Mortality

*Anthony B. Miller, Teresa To, Cornelia J. Baines, Claus Wall**

The Canadian National Breast Screening Study (CNBSS), conducted on women age 40–49, was designed to evaluate the efficacy of combined annual mammography and physical examination of the breasts in reducing breast cancer mortality in comparison to usual care (UC) controls. From January 1980 through March 1985, 25,214 women were individually randomized to the mammography/physical exam (MP) arm and 25,216 to the UC. The integrity of the randomization has been reviewed and confirmed to be unbiased. During an average follow-up of 10.5 years from entry (range: 8.75–13 years), 82 women died from breast cancer in the MP arm and 72 in the UC, for a rate ratio of 1.14 (95% confidence interval: 0.83–1.56). All-cause mortality was almost identical comparing the two groups; the nonsignificant excess of breast cancer deaths in the MP arm was balanced by an excess of other cancer deaths in the UC arm. [Monogr Natl Cancer Inst 1997:22:37–41]

The Canadian National Breast Screening Study (CNBSS) is an individually randomized trial designed to evaluate, in women age 40–49 on entry to the study, the combined efficacy of annual mammography, physical examination of the breasts, and the teaching of breast self-examination in reducing breast cancer mortality (*1*). Thus, it was specifically designed to evaluate the *efficacy* of screening in women who chose to be screened, rather than evaluating the *effectiveness* of screening in the population. Efficacy trials are usually regarded as necessary before effectiveness (population-based) trials are conducted. To date, it is the only trial specifically designed to evaluate screening in women age 40–49, rather than in a wider age range, that has reported upon breast cancer mortality.

In our published seven-year mortality report (*2*), we demonstrated that the two arms of the study were well balanced with respect to age, marital status, number of live births, menopausal status, education, family history of breast cancer, and place of birth. The validity of the randomization has since been challenged (*3*). More women with breast cancer with four or more nodes were identified at the initial screening examination in the treatment, or mammography/physical exam (MP) arm, than in the usual care (UC) control arm. However, Bailar and MacMahon (*4*) carried out an independent review of randomization for the National Cancer Institute of Canada, paying particular attention to the centers where the excess was concentrated, and they found no evidence of any deliberate falsification of randomization such that more women with ''advanced'' breast cancers were placed in the MP arm. Further, an independent validation of CNBSS data from the Manitoba screening center has found no evidence of falsification there either (*5*). A commentary by Boyd (*6*) attempted to cast some doubt on whether, ''the debate is over.'' Accordingly, we shall try to put the record straight in what follows.

The other issue that has surfaced relates to mammography quality (*7–9*). Several procedures were put in place in the CNBSS to obtain high-quality mammography. Centers with mammography experience were selected, dedicated mammography machines and film processing were required, modern film-screen technology was used, there was extensive reference physicist (*10*) and reference radiologist (*11,12*) monitoring, external reviews of mammography were conducted (*13,14*), and the findings were reported back to the study centers. Our procedures were designed to maximize the sensitivity of the screen, even at the cost of reduced specificity. As Fletcher et al. (*15*) have reported, these efforts resulted in parameters of quality that rivaled all other screening trials in this age group. As a result, the sensitivity of the screen in the MP arm was 81%, the first round breast cancer detection rate was 3.9 per 1,000, the prevalence/incidence ratio was 2.7, and 65% of the invasive screen-detected breast cancers were node negative.

Methods

Women with no previous history of breast cancer and no mammogram in the previous 12 months were eligible for the trial, providing they signed an informed consent form. A total of 50,430 women age 40–49 were enrolled from January 1980 through March 1985 from 15 centers across Canada. Randomization was to mammography and physical examination of the breasts (the MP allocation) or to a control group receiving usual care in the context of the Canadian health care system (the UC allocation). Randomization was performed by the local coordinators by reference to prearranged lists, after the coordinators had received from the examiner the signed informed consent and completed initial physical examination forms. This was to ensure that the physical examination would be conducted and the findings recorded without knowledge as to whether mammography was allocated. In the MP allocation, five annual screens were offered to the majority of participants; those enrolled in the last year of recruitment in the individual centers were only offered four annual screens. The participants in the UC arm received annual questionnaires over the same time period. Com-

**Affiliations of authors:* National Breast Screening Study, Department of Public Health Sciences, University of Toronto, Canada.

Correspondence to: Anthony B. Miller, Department of Public Health Sciences, Faculty of Medicine, University of Toronto, Toronto, Ontario, M5S 1A8, Canada.

See ''Notes'' following ''References.''

pliance with rescreening and with returning questionnaires was excellent, exceeding 90% in both arms. Breast cancer mortality has been ascertained by annual follow-up of all women known to have been diagnosed with breast cancer, and by linking CNBSS records to the Canadian National Mortality Data Base (CNMDB), initially for deaths up to December 31, 1988, and more recently to December 31, 1993, the closing date for the present analysis. Thus, participants have been followed for a mean of 10.5 years, with a range of 8.75 to 13 years.

The trial was planned with sufficient power to detect a 40% reduction in breast cancer mortality at five years from entry (*1*). However, because insufficient deaths from breast cancer had occurred by five years to attain the planned power, the follow-up was extended for two years, by which time there were enough breast cancer deaths to reach the planned power (*2*). The present report provides the findings from an additional three years of follow-up, providing sufficient power to detect at least a 30% reduction in breast cancer mortality.

Results

If women had been deliberately placed in the MP arm because of concern over their possible risk of breast cancer—due to, say, a strong family history of breast cancer—an excess of women with risk factors would have been detected in the MP arm. As shown in Table 1, that was not so. If, on the other hand, the concern was that the woman already had signs or symptoms of breast cancer, the examiner would have identified an abnormality. However, all women with clinically detected abnormalities were referred to the CNBSS review clinic to be assessed by the study surgeon. Table 2 demonstrates that such referrals were similar across the two arms within the study centers. Other analyses have shown women who reported breast symptomatology at the time of their initial physical examination were equally distributed.

Using the data from the CNMDB to December 31, 1993, we identified 82 breast cancer deaths in the MP arm and 72 in the UC arm. In terms of person-years of observation to December 31, 1993, this is a rate ratio of 1.14 (95% confidence interval [CI]: 0.83–1.56). The cumulative mortality from breast cancer over the 13 years of observation is presented in Figure 1.

The distribution of breast cancer deaths in relation to various factors is presented in Table 3. Although there are some inequalities by five-year age groups, the differences are not statistically significant. Less than half the breast cancer deaths in both arms were among women referred to review after the first screen. Again, there are no differences between the arms. We have noted elsewhere that the detection of an abnormality on physical examination was a risk factor for subsequent breast cancer detection, as it was for mammography (*16*). A minority of breast cancer deaths were among women with breast cancers detected at the first screen. There were more in the MP than the UC arm, as a result of the additional cancers found by mammography. When all breast cancer deaths were related to nodal status at the time of diagnosis of the cancer, we found a higher proportion of deaths among women with cancers labeled as node negative in the UC arm.

Table 4 shows the distribution of deaths from all causes up to December 31, 1993. Although breast cancer was the largest

Table 1. Distribution of epidemiologic variables from initial questionnaire, women age 40–49

Factor	MP	UC
Marital status:		
Never married	6.5%	6.5%
Married now	80.6%	80.7%
Live births:		
None	15.0%	15.3%
1–3	66.6%	65.8%
4+	18.4%	19.0%
Menopausal status:		
Premenopausal	66.4%	67.1%
Postmenopausal	4.9%	4.8%
Hysterectomy	25.6%	25.1%
Education:		
Grade school	44.9%	45.2%
Higher	55.1%	54.8%
Family history of breast cancer:		
Mother	8.1%	8.2%
Sister	3.3%	3.5%
Second degree	26.2%	26.7%
Place of birth:		
North America	84.3%	84.3%
Europe	13.2%	13.3%

Table 2. Women age 40–49 on entry with abnormalities detected on clinical examination at initial screen and referred to review

Screening center	MP	UC
1. Mount Sinai Hospital, Toronto	420	488
2. Saint Sacrement Hospital, Quebec	643	675
3. Notre Dame Hospital, Montreal	458	469
4. Henderson General Hospital, Hamilton	218	243
5. Health Science Center, Winnipeg	265	231
6. Cancer Center, Vancouver	399	371
7. Ottawa Civic Hospital	85	89
8. Ottawa General Hospital	73	68
9. Hôtel Dieu Hospital, Montreal	128	179
10. Halifax General Hospital	237	214
11. Westminster Hospital, London	167	162
12. Cross Cancer Institute, Edmonton	129	136
13. St. Michael's Hospital, Toronto	39	50
14. Red Deer General Hospital	58	42
15. Tom Baker Cancer Center, Calgary	179	184
Total, all centers	3498	3601

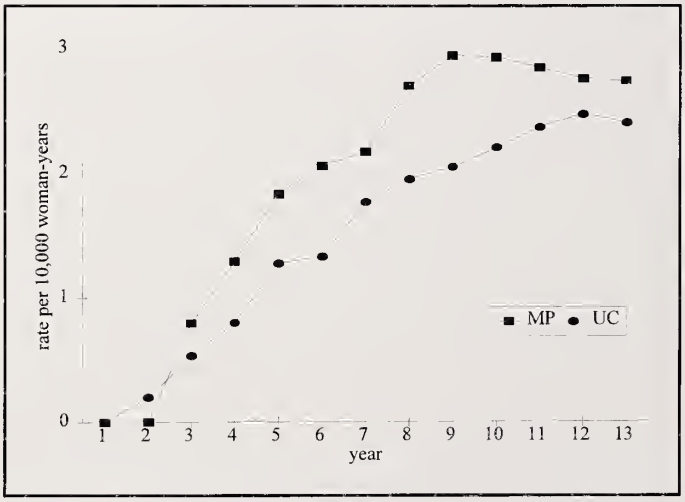

Fig. 1. Cumulative breast cancer mortality in the MP and UC allocations.

Table 3. Deaths due to breast cancer, women age 40–49 on entry

Factor	MP Number	MP Percent	UC Number	UC Percent
Age at entry:				
40–44	32	39	38	53
45–49	50	61	34	47
Referred to review, screen 1				
Yes	34	41	27	38
No	48	59	45	62
Nodal status, screen 1 cancers				
Negative	3	4	6	8
1–3 nodes	4	5	3	4
4+ nodes	8	10	1	1
Unknown	0	0	0	0
Subsequent detection	67	82	62	86
Nodal status, all cancers				
Negative	23	28	34	47
1–3 nodes	18	22	9	12
4+ nodes	27	33	16	22
Unknown	14	17	13	18
Total	82	100	72	100

Table 4. Underlying cause of death by allocation, women age 40–49 on entry

ICD CODE	MP Number	MP Percent	UC Number	UC Percent
Breast cancer	82	19.6	72	17.4
Lung cancer	42	10.0	42	10.1
Colorectal cancer	23	5.5	32	7.7
Stomach cancer	5	1.2	12	2.9
Pancreas cancer	17	4.1	14	3.4
Uterus/Cervical cancer	5	1.2	7	1.7
Ovary cancer	22	5.3	21	5.1
Hematopoietic neoplasm	28	6.7	25	6.0
Other neoplasms	57	13.6	58	14.0
Central nervous system	13	3.1	8	1.9
Circulatory	54	12.9	43	10.4
Respiratory	10	2.4	13	3.1
External	32	7.7	35	8.5
Other causes	28	6.7	32	7.7
All causes	418	100.0	414	100.0

single cause of death, it accounted for only 19.6 % of the deaths in the MP arm and 17.4% in the UC. There are minor differences in some categories, but in general, the reported causes of death were remarkably similar, thus providing further confirmation that the randomization resulted in comparable groups.

Discussion

The present report more than doubles the number of breast cancer deaths previously noted at seven years. In the seven-year report (*2*), there were 38 deaths from breast cancer in the MP and 28 in the UC allocation. The ratio of the proportions of breast cancer deaths in the MP allocation compared to the UC was 1.36 (95% CI: 0.84–2.21). Breast cancer mortality figures derived from our routine annual follow-up of all the breast cancers ascertained in the study were included in the summary report from the March 1996 meeting in Falun, Sweden, resulting in 78 in the MP arm and 73 in the UC (*17*). Currently, at a mean follow-up time of 10.5 years, we are able to exclude, with 95% confidence, a reduction of breast cancer mortality of 17% or more. Although the absolute level of the nonsignificant excess of breast cancer mortality found previously in the MP arm has not changed comparing seven year to current results, proportionately it is now much less.

Having failed to find a benefit from screening in women who initiate screening at ages 40–49, the CNBSS has been subjected to intense review and criticism. Similar scrutiny has not been applied to trials that did report a benefit. For example, in his comments on the meta-analysis by Smart et al. (*18*), Boyd (*6*) fails to note that the trials, other than HIP, that contributed to the suggestion "that mammography is effective in reducing the rate of death from breast cancer in this age group" (*18*) have not published data confirming equivalence of subjects in the compared arms at the time of randomization, as has the CNBSS. Indeed, many are cluster-randomized trials, and differences between the clusters are to be expected; yet, the design effect of the cluster randomization has not been factored into the meta-analysis, so that the confidence intervals reported are too narrow.

A great deal of attention has also been given to the excess of breast cancers with four or more positive nodes in the first screen in the MP arm compared to the UC in the CNBSS (*3,6*). Variables that become apparent as a result of screening and diagnosis have been called pseudo-variables by Prorok et al. (*19*) and are biased. Nodal status is one such variable. Whether the CNBSS study surgeon referred a woman with a physical "abnormality" for subsequent diagnosis (and biopsy) in the community was influenced by the availability of mammograms in the MP group and their nonavailability in the UC group. Several women with four or more nodes were probably unrecognized in the UC group, and many were not even recognized as node positive subsequently, as they were more likely to be treated in centers where careful extensive nodal dissection or evaluation by skilled pathologists was not the norm. Some may not have had nodal dissection at all. Moreover, in the MP arm, many of the so-called "advanced" cancers were small, with limited involvement of the individual nodes, and were thus not clinically advanced, even though four or more nodes were found to be involved after careful dissection and histologic sectioning.

The higher proportion of breast cancer deaths among node-negative women in the UC arm is further evidence that the difference in nodal status between the MP and UC groups detected at initial screening was partly due to failure to identify as node positive a number of the breast cancers in the UC arm. That the initial excess of four or more node positive cancers in the MP arm is due to a diagnostic bias is confirmed by the similarity in numbers of breast cancer deaths among women with cancers ascertained either by screening or as interval cancers in the first 12 months after entry (see Baines, this volume). An explanation for the persistent excess of 10 breast cancer deaths in the MP arm may be found in Table 4, which shows a deficit of colorectal and stomach cancer deaths. It seems possible that this is an example of the "sticking diagnosis" phenomenon. Women diagnosed with breast cancer as a result of mammography screening, and who developed metastatic disease, may be less likely to be investigated for a new primary tumor than women without a breast cancer diagnosis in the UC arm. Thus, it is possible that some of the breast cancer deaths in the MP arm were in fact due to a second primary in the gastrointestinal tract.

The initial physical examinations in the CNBSS have been

referred to as a "prescreen" by some commentators. That is incorrect. The physical examinations were administered as screening tests that were evaluated and subject to quality control in the same way as mammography. Both groups were initially screened by physical examination, an approach in the UC group that mimics what a careful physician might be expected to perform on women in this age group in North America before deciding whether to prescribe mammography. About a quarter of the women in the UC arm received one or more mammograms during the course of the trial, as was expected from the ready availability of mammography in the Canadian health care system. That was not "contamination"; it was good usual care. We have demonstrated that substituting annual mammography and physical examinations for such usual care during a four-year period has no impact on breast cancer mortality over a 8.75- to 13-year period.

"Modern" mammography is said to be much improved compared to CNBSS mammography. But what is the nature of the improvement? Few data have been presented that support increased sensitivity from the mammography of the 1990s compared to that of the 1980s. Rather, what has happened is a major improvement in specificity, reducing anxiety in screened women and health care costs, but having no impact upon breast cancer mortality.

Boyd (*6*) and others have commented that longer follow-up of the existing trials over the next few years "should settle the debate." This seems unlikely, given the lack of any indication of benefit with longer follow-up in the CNBSS. Further, the lead time gained by the MP screen in the CNBSS in women age 40–49 was 2.3 years (95% CI: 1.5–3.2) compared to 3.6 years (95% CI: 2.7–5.5) for women age 50–59 (To T, Miller AB, Xie HX, Walter S., "Lead time estimation and its use in survival analyses as applied to the National Breast Screening Study," submitted, 1997). This supports other studies that suggest that the rate of progression of breast cancer in premenopausal women is faster than in postmenopausal women (*17*). This makes it unlikely that a delayed benefit of breast cancer screening in younger compared to older women explains the trends seen after 10 years in some studies of the long-term follow-up of women screened initially under and over the age of 50 (*20*), in spite of attempts to provide a rationale for this paradoxical finding (*17*).

One reason for the CNBSS not showing a breast cancer mortality reduction (even though some other trials have suggested a benefit beginning after seven years from entry) may be the smaller size of the tumors in the control arm of the CNBSS compared to control women in the Swedish Two-County Trial (Narod S, "On being the right size: a reappraisal of mammography trials in Canada and Sweden," submitted, 1997). This would explain the superior survival of UC women with breast cancer at seven years (*2*) compared to those in the Swedish Two-County Trial (*21*). Further, there was almost universal use of adjuvant chemotherapy for node-positive breast caneer in Canada during the 1980s, whereas adjuvant chemotherapy was not used in the trials in Sweden that began in the 1970s (Tabar, L, personal communication, 1997). It has been suggested, on the basis of the Two-County Trial, that only a limited proportion of breast cancers can benefit from early detection from screening (*17*). If this segment is benefited by usual care in the Canadian health care context, or is the same as can be cured by adjuvant chemotherapy, it is scarcely surprising that screening cannot be shown to make an additional impact.

In the light of our results, what should women be advised? It seems important that women should understand that the largest trial to date shows no evidence of benefit from initiating mammography screening under the age of 50. This negative finding, however, must be placed in the context that the CNBSS is the only trial since HIP designed specifically to evaluate screening in North America. Still, even two-view mammography, conducted annually, has not resulted in the earlier detection of curable cancers which would be fatal in the absence of their early detection. Thus, although women may choose to be screened by mammography, they should understand that usual care, as defined in Canada with the ready availability of physical examinations of the breasts, the practice of breast self-examination, diagnostic mammography, and good cancer treatment, seems an extremely viable option.

In closing, we emphasize that this is a preliminary update from our recent linkage between the CNBSS file and the Canadian National Mortality Data Base. There will probably be some minor changes in the numbers reported in this paper, as the breast cancer deaths now known to us are not the final tally for the 10- to 15-year report currently in preparation. Only when we are able to evaluate the findings from the record linkage to the Canadian National Cancer Registry, currently underway, will we be able to produce the final tally. Nevertheless, it seems unlikely that our present findings will change to any great degree.

Conclusion

The CNBSS is internally valid and there is no evidence of bias in allocation. Screening of women age 40–49 with yearly mammography and physical examination has had no impact on mortality from breast cancer during 8.75 to 13 years from entry.

References

(*1*) Miller AB, Howe GR, Wall C. The National Study of Breast Cancer Screening: protocol for a Canadian randomized controlled trial of screening for breast cancer in women. Clin Invest Med 1981;4:227–58.

(*2*) Miller AB, Baines CJ, To T, Wall C, et al. Canadian national breast screening study: 1. Breast cancer detection and death rates among women age 40–49 years. Can Med Assoc J 1992;147:1459–76.

(*3*) Tarone RE. The excess of patients with advanced breast cancer in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 1995;75:997–1003.

(*4*) Bailar JC, MacMahon B. Randomization in the Canadian National Breast Screening Study. Report of a review team appointed by the National Cancer Institute of Canada. Can Med Ass J 1997;156:213–5.

(*5*) Cohen MA, Kaufert PA, MacWilliam L, Tate RB. Using an alternative data source to examine randomization in the Canadian National Breast Screening Study. J Clin Epidemiol 1996;49:1039–44.

(*6*) Boyd NF. The review of randomization in the Canadian National Breast Screening Study. Is the debate over? Can Med Ass J 1997;156:207–9.

(*7*) Kopans DB, Feig S. The Canadian National Breast Screening Study: a critical review. AJR Am J Roentgenol 1993;161:755–60.

(*8*) Burhenne LJ, Burhenne HJ. The Canadian National Breast Screening Study: a Canadian Critique. AJR Am J Roentgenol 1993;161:761–3.

(*9*) Boyd NF, Jong RA, Yaffe MJ et al. A critical appraisal of the Canadian National Breast Screening Study. Radiology 1993;189:661–3.

(*10*) Yaffe MJ, Mawdsley GE, Nishikawa RM. Quality assurance in a National Breast Screening Study. SPIE 1983;419:23–30.

(11) Baines CJ, McFarlane DV, Wall C. Audit procedures in The National Breast Screening Study: mammography interpretation. J Can Assoc Rad 1986;39:256–60.
(12) Baines CJ, McFarlane DV, Miller AB. The role of the reference radiologist: estimates of inter-observer agreement and potential delay in cancer detection in the National Breast Screening Study. Investigative Radiology 1990; 25:971–6.
(13) Baines CJ, Miller AB, Kopans DB, et al. Canadian National Breast Screening Study: assessment of technical quality by external review. AJR Am J Roentgenol 1990:155:743–7.
(14) Miller AB, Baines CJ, Sickles EA. Canadian National Breast Screening Study. AJR Am J Roentgenol 1990;155:1133–4.
(15) Fletcher SW, Black W, Harris R, et al. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993,85: 1644–56.
(16) Holowaty PH, Miller AB, Baines CJ, Risch H. Canadian National Breast Screening Study: first screen results as predictors of future breast cancer risk. Cancer Epidemiology, Biomarkers & Prevention 1993;2: 11–9.
(17) Report of the Organizing Committee and Collaborators. Breast cancer screening with mammography in women aged 40–49 years. Int J Cancer 1996;68:693–9.
(18) Smart CR, Hendrick RE, Rutledge JH III, Smith RA. Benefit of mammography screening in women age 40–49: current evidence from randomized controlled trials. Cancer 1995;75:1619–26.
(19) Prorok PC, Hankey BF, Bundy BN. Concepts and problems in the evaluation of screening programs. J Ch Dis 1981;34:159–71.
(20) Kerlikowske K, Grady D, Rubin SM, et al. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.
(21) Tabar L, Fagerberg G, Chen HH, Duffy SW. Screening for breast cancer in women aged under 50: mode of detection, incidence, fatality and histology. J Med Screening 1995;2:94–8.

Notes

We acknowledge the critical contributions, during the conduct of the CNBSS, of the reference radiologist, the late Dr. DV McFarlane, the reference physicist, Dr. MJ Yaffe, and the following CNBSS clinical investigators:

Center Directors: AA Bassett, DCG Bethune, DM Bowman, H Bush, J Cantin, L Dêschenes, JE Devitt, DN Graham, G Hislop, AW Lees, BM Lefébvre, L Mahoney, SE O'Brien, A Simard, WJ Temple.

Surgeons: CP Armstrong, RM Baird, the late W Beecroft, WJ Buie, R Bury, CDJ Chadwick, WG Chipperfield, D Currie, GJ Dewar, M Falardeau, GJ Francis, MH Friedman, N Gagic, D Girvin, HR Harse, DJ Hamilton, I Koven, U Kuusk, RD Marriott, AB McCarten, J McCredie, WO Onerheim, A Péloquin, C Potvin, RE Pow, J Purves, PM Rebbeck, J Robert, A Robidoux, JT Sandy, S Sidlofsky, ER Sigurdson, B Steele, RM Stone, JB Taillefer, TK Thorlakson, GK Thorson.

Radiologists: L Audet, BL Bird, MJ Burns, B Capusten, WR Castor, GM Cooke, CM Copeland, JW Davidson, GED Davis, JEL Desautels, RL Desmarais, LA Fried, A Grégoire, G Hardy, P Hassell, G Hébert, R Jong, SM Kelly, the late J Ladouceur, J Laperrière, JD Longley, RN Ludwig, JHM MacGregor, J McCallum, JS Manchester, HF Morrish, HA Mueller, T Minuk, D Ouimet-Oliva, P Poon, O Prosmanne, P Rasuli, NL Patt, M Petitclerc, JW Radomsky, JL Robillard, the late IS Simor, RK Sparrow, BJ Shapiro, SL Share, HK Standing, WJ Weiser, AH Zalev.

Pathologists: F Alexander, Y Boivin, N Cooter, J Danyluk, D Dawson, TJ D'Souza, M Jabi, S Jacob, J Safneck, W Schurch, H Strawbridge, DI Turnbull, R Vauclair, A Worth, H Yazdi, I Zayid.

The CNBSS was supported by the Canadian Breast Cancer Research Initiative, the Canadian Cancer Society, the Department of National Health and Welfare, the National Cancer Institute of Canada, the Alberta Heritage Fund for Medical Research, the Manitoba Health Services Commission, the Medical Research Council of Canada, le Ministère de la Santé et des Services Soçiaux du Québec, the Nova Scotia Department of Health, and the Ontario Ministry of Health.

Recent Results From the Swedish Two-County Trial: The Effects of Age, Histologic Type, and Mode of Detection on the Efficacy of Breast Cancer Screening

*László Tabár, Hsiu-Hsi Chen, Gunnar Fagerberg, Stephen W. Duffy, Teresa C. Smith**

The effect of mammographic screening in reducing mortality from breast cancer is known to be smaller and more delayed in women aged 40–49 than in women over 50. In this study, we investigated how these phenomena relate to histology-specific breast cancer incidence and mortality. The data are from 2,468 women with breast cancer who participated in the Swedish Two-County Trial. The overall relative breast cancer mortality of invited to noninvited women aged 40–49 was 0.87, and the relative mortality from poorly differentiated (grade 3) ductal carcinoma was 0.95. These results were not statistically significant. The corresponding relative risks for invited women aged 50–74 were a statistically significant 0.65 and 0.61. We conclude that in this trial, with a two-year interscreening interval, the smaller and later effect of invitation to screening on breast cancer mortality in women 40–49 years old is due to the failure of screening to reduce mortality from grade 3 ductal carcinoma in this age group. [Monogr Natl Cancer Inst 1997;22:43–47]

Two main tumor characteristics seem to play a crucial role in controlling breast cancer: heterogeneity of the disease and its progressive nature (*1,2,3,4*). Because mammography screening can allow earlier diagnosis and treatment of breast cancer, it can significantly decrease Stage II and more advanced tumors. Since an advanced disease stage is strongly associated with death from breast cancer, the relative incidence rate of Stage II and more advanced cases among women invited to screening compared to those not invited is expected to be a sensitive measure of breast cancer mortality. The close correlation between the cumulative incidence rates of advanced breast cancers and cumulative mortality rates has been well documented in different screening trials (*5,6,7,8*).

The relationship between advanced breast cancer rates and disease-specific mortality rates has also been demonstrated in age subgroups (*9*). The relative incidence of tumors Stage II and higher is consistent with the diminished effect of screening on mortality in women aged 40–49 years, but it does not explain the reason for the delayed benefit in this age group. Also, it raises the question of why there is less reduction in advanced tumors and subsequently in breast cancer mortality in women aged 40–49 compared to older women. Investigating the heterogeneity of breast cancer, comparing the impact of mammographic screening on cancers of different histologic types, and analyzing the variability in tumor progression rates by age may give insight into the varying efficacy of screening in different age groups.

Survival analysis based on the Swedish Two-County Trial confirms that breast cancer cases can be classified into three histologic tumor types according to prognosis: Group I (consisting of ductal carcinoma *in situ* [DCIS], grade 1 invasive ductal carcinomas, tubular cancers, and mucinous cancers) has good survival, Group II (grade 2 invasive ductal, medullary, and invasive lobular cancers) has intermediate survival, while Group III (grade 3 invasive ductal cancers) has poor survival (*10*).

In previous studies, we concluded that the duration of the tumors' preclinical detectable phase (sojourn time), and therefore the rate of progression from the preclinical to the clinical phase, varies considerably not only by histologic type but also by patient age (*11*). The practical implication of this is that the impact of screening on mortality from breast cancer, and the timing of this impact, will depend largely on which histologic types will be diagnosed early in their natural history and whether screening will advance the time of diagnosis of the subgroup with poor prognosis.

The poorly differentiated invasive ductal carcinomas that make up Group III have both a rapid progression from the preclinical to the clinical phase (a short sojourn time) and a poor short-term survival. Therefore, early detection of these high-risk cases will have a demonstrable beneficial effect within a few years following diagnosis and treatment (short-term effect). On the other hand, the impact of early detection of tumors in Group I and Group II on mortality from breast cancer will not be demonstrable until many years later (long-term effect), since women with similar but undetected tumors in the control group will live much longer than those with poorly differentiated tumors.

As we have noted previously, the relative mortality invited to noninvited women in the Two-County Study was 0.87 in the 40–49 age group and 0.65 in the 50–74 group (*11*). Since the tumor progression rate from preclinical to clinical phase is more

**Affiliations of authors:* L. Tabár, Department of Mammography, Central Hospital, 79 182 Falun, Sweden; H.-H. Chen, S.W. Duffy, T.C. Smith, MRC Biostatistics Unit, Cambridge, UK; G. Fagerberg, Department of Mammography, University of Linkoping, Sweden.

Correspondence to: László Tabár, Department of Mammography, Central Hospital, 79182 Falun, Sweden (Tel. 46 23 82507)

rapid in younger than in older women (*11,12,13*), the smaller benefit of mammography screening for women under 50 in the Two-County Trial might be due to the longer interscreening interval, which did not allow sufficiently early detection of rapidly growing and frequently fatal tumors, such as poorly differentiated grade 3 ductal carcinomas. Analysis of the cumulative incidence rate of Stage II and worse cancers by histologic type and age will test this hypothesis.

The purpose of this study, then, is to:

(1) consider whether the effect of invitation to mammography screening on mortality from breast cancer is uniform for all tumor types, or if the reduction in mortality is more pronounced for some histologic types;
(2) examine whether the impact of screening on mortality from different histologic tumor types varies with age;
(3) compare the cumulative incidence rate of Stage II and more advanced (Stage II+) tumors with the corresponding observed mortality in each histologic group; and
(4) make suggestions for mammography screening of women aged 40–49 years, based on (1), (2) and (3).

Methods

Data Source

Data used in this study are from 2,468 women diagnosed with breast cancer who participated in the Swedish Two-County Trial: 1,053 and 1,415 were from the W and E counties respectively. Average follow-up was 14 years through December 31, 1994. Screening intervals for the invited groups were 24 and 33 months, respectively, for women aged 40–49 and those over 50. (Note that although we refer to the younger age group as the "40–49" group, 30% of follow-up screens in this age group actually took place after the women had reached age 50.) The prospectively determined histologic tumor types include ductal carcinoma *in situ* (DCIS), invasive ductal carcinomas not otherwise specified (NOS) of malignancy grades 1, 2 and 3, and medullary, invasive lobular, tubular, and mucinous carcinomas. Details of the study design have been described fully elsewhere (*6*). Note that in this paper, we follow the convention, employed whenever reporting results of the Two-County Trial, of referring to the group invited to screening as the Active Study Population (ASP) and the uninvited control group as the Passive Study Population (PSP).

Statistical Methods

Cumulative mortality rates were calculated by dividing deaths from breast cancer of various histologic types by person-years.

Calculation of relative risk of cumulative incidence of Stage II+ or cumulative mortality since time at entry is by Poisson regression analysis (*14*).

Results

Table 1 shows the cumulative mortality by tumor type for the ASP and PSP. Statistically significant reductions of 37% and 39%, respectively, can be seen in deaths from grade 2 and grade 3 invasive ductal carcinomas in invited women aged 50–74 at randomization. In the 40–49 group, most of the mortality reduction is confined to Group II tumors (grade 2 invasive duetal cancers, medullary cancers, and invasive lobular carcinomas), and a 5% reduction in death from grade 3 invasive ductal carcinoma was observed. The absolute risk of dying from breast cancer in Group I (DCIS and grade 1 ductal, tubular, and mucinous carcinomas) is negligible in comparison to deaths from breast cancers in Groups II and III, although the relative risk is high due to the large number of Group I cancers detected at screening.

As noted above, detecting tumors of various histologic types at an earlier stage will be expected to have varying effects on the short-term and long-term mortality results. Early detection of high-risk (Group III) breast cancer cases will reduce mortality a few years after randomization, since poorly differentiated clinically diagnosed cancers are often associated with poor short-term survival (within five years). The beneficial effect of early detection of intermediate risk (Group II) cancers will not be

Table 1. Cumulative mortality (number of deaths) per 100,000 from breast cancers by histologieal tumor type and age in women invited to screen (ASP) and not invited to screen (PSP), with relative risk (RR) of breast cancer death, Swedish Two-County Trial

	40–49			50–74		
Age group histology	PSP (PY* = 226,526)	ASP (PY = 278,703)	RR (95% CI)	PSP (PY = 543,939)	ASP (PY = 772,979)	RR (95%) CI
Grade 3 ductal	10.59	10.05	0.95	23.90	14.23	0.61
(Group III)	(24)	(28)	(0.55–1.64)	(130)	(110)	(0.47–0.78)
Grade 2 ductal	3.09	2.15	0.70	12.31	7.76	0.63
(Group II)	(7)	(6)	(0.23–2.07)	(67)	(60)	(0.44–0.89)
Lobular	1.77	1.43	0.81	4.60	2.85	0.62
(Group II)	(4)	(4)	(0.20–3.25)	(25)	(22)	(0.35–1.10)
Medullary	1.32	0.36	0.27	0.92	0.52	0.56
(Group II)	(3)	(1)	(0.03–2.60)	(5)	(4)	(0.15–2.10)
Grade 1 ductal	0	1.43	—	1.84	1.55	0.84
(Group I)		(4)		(10)	(12)	(0.36–1.95)
Mucinous	0	0	—	0.55	1.29	2.34
(Group I)				(3)	(10)	(0.65–8.52)
Tubular	0	0	—	0	0.39	—
(Group I)					(3)	
DCIS	0	1		0.18	0.26	1.41
(Group I)				(1)	(2)	(0.13–15.52)

*PY: Person-years

demonstrable until around eight years after randomization, when those tumors with intermediate survival will result in breast cancer death in the control group.

Our results support this varying effect, as shown in Figures 1–4. The reduction in mortality from grade 3 ductal carcinoma begins to appear four to five years after randomization in the 50–74 age group (Figure 1b) and is hardly apparent in the 40–49 age group (Figure 1a). The diminished impact of grade 3 ductal cancers on mortality in the 40–49 age group explains both the reduced effect of screening on breast cancer death in younger women and the lack of short-term benefit. The deaths prevented in this age group were from the Group II cancers (grade 2 invasive ductal, medullary, and invasive lobular), showing a demonstrable benefit only after six to eight years in both age groups (Figures 4a and 4b).

We also compared the cumulative mortality by histologic type and age (as shown in Figures 1 and 2) with the cumulative incidence rates of Stage II+ cancers by histologic type and age. The reductions in mortality from Group III cancers were 5% and 39% for 40–49 and 50–74 age groups respectively (Figure 1). The corresponding reductions in Group III cancers of Stage II or worse were 0% and 37%. The mortality reductions from Group II tumors were 36% in both age groups (Figure 2). The reduction in Stage II+ tumors were 28% and 35%. These findings suggest that two-year screening in the 40–49 age group failed to detect grade 3 tumors at an early stage, which in turn resulted in similar incidence of Stage II+ cancers in both the invited and control groups. This indicates that poorly differentiated ductal cancers have a more rapid progression during their preclinical phase in women aged under 50 compared to women 50 years of age and older.

Discussion

Our analysis of mortality and incidence of Stage II+ breast cancer cases according to histologic tumor type has demonstrated a considerable reduction in mortality from poorly differentiated ductal breast carcinomas in women aged 50–74 years at randomization, in spite of the long 33-month interscreening interval, while only a 5% reduction was achieved with the 24-month interval in women aged 40–49 years. These findings imply that the aggressive tumors are more amenable to early detection when they occur at a later age in the host's life.

The more rapid progression of grade 3 ductal cancers in younger women makes early detection more difficult. This is reflected in steady incidence of Stage II+ grade 3 ductal cancers in the younger age group. The lesser and later mortality reduction in women aged 40–49 years can be explained by the fact that the mortality reduction is limited to the histologic types with intermediate survival—that is, to grade 2 ductal, medullary, and invasive lobular cancers.

Our results may also explain the difference in results observed between the two counties. We have published a 27% reduction in mortality from breast cancer in the 40–49 age group in W-county as opposed to a 0% reduction in E-county (*2*). When plotting the cumulative mortality curves by histologic type and age in W-county, we observed a mortality reduction from grade 3 ductal cancers both under and over age 50, although there was a lesser reduction in the younger age group (Figure 3). The reduction in E-county was confined to the age group 50–74 (Figure 4). It should be kept in mind, however, that in the invited group in E-county, age 40–49 at randomization, five breast cancer deaths occurred among nonattenders with grade 3 cancer (*10*).

Several factors have contributed to the decrease in mortality from breast cancer in the Two-County Trial. One of the most important is reducing the tumor size and frequency of axillary lymph node metastases of grade 3 invasive ductal carcinomas. Also, there is a 15% reduction in the incidence of poorly differentiated ductal cancers in the ASP compared to the PSP in age group 50–74 years, suggesting that some of the tumors may dedifferentiate during growth and that early detection may stop progression of the malignancy grade. This is consistent with our earlier findings, according to which approximately 50% of grade 1 and 2 ductal cancers have the potential to dedifferentiate during growth in women aged 50–69 years (*11*). We have not found a reduction in the incidence of grade 3 ductal cancers in women aged 40–49 years. This suggests that dedifferentiation of grade 1 and 2 cancers occurs rapidly in this age group, during the short preclinical phase (*12*), and can only be prevented by shortening the interscreening interval.

The length of the interscreening interval for women aged

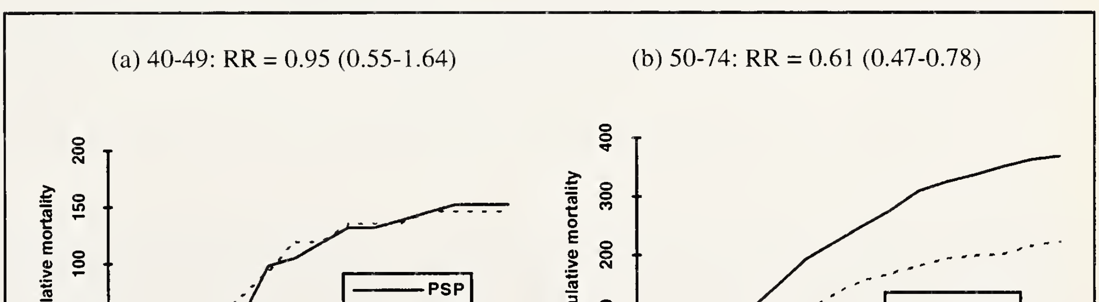

Fig. 1. Cumulative mortality from breast cancer for ductal-grade 3 carcinoma by age, Swedish Two-County Trial.

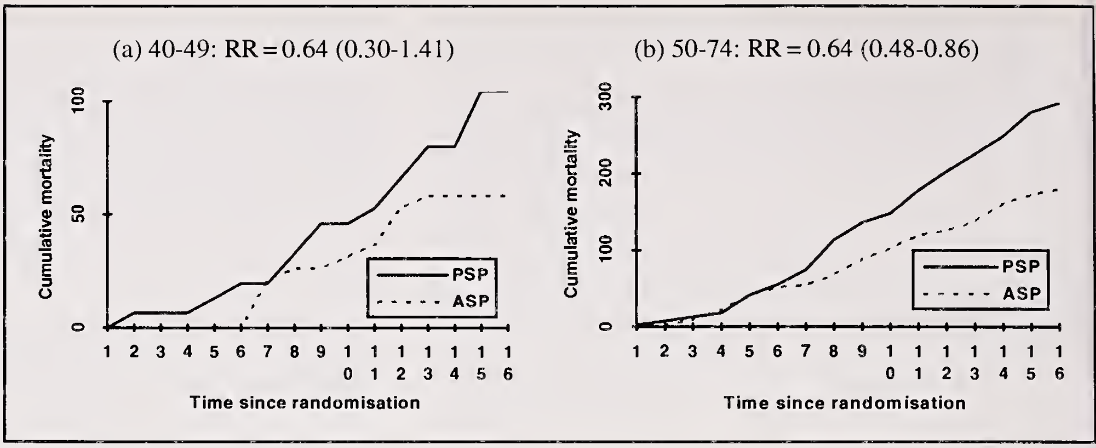

Fig. 2. Cumulative mortality from breast cancer for ductal-grade 2, lobular and medullary carcinoma by age, Swedish Two-County Trial.

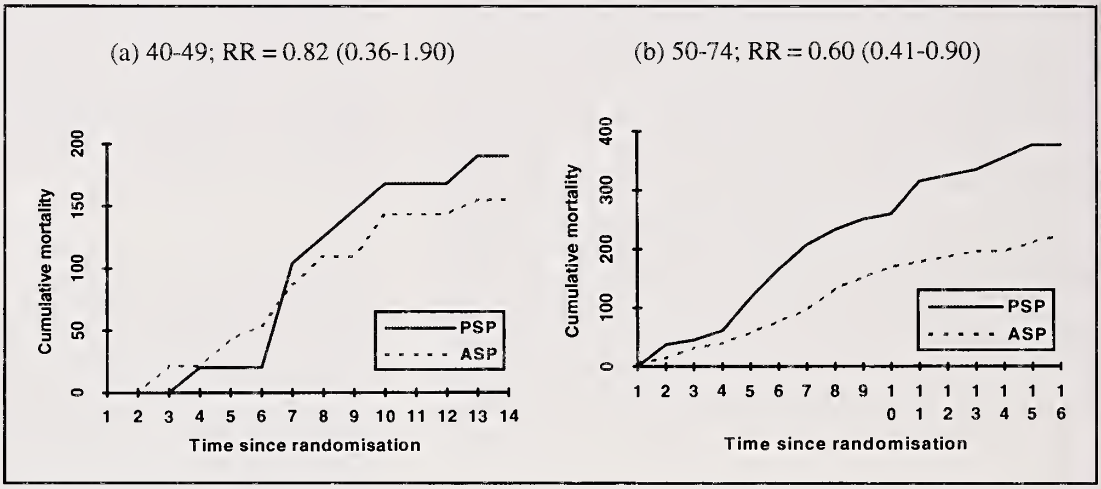

Fig. 3. Cumulative mortality from breast cancer for ductal-grade 3 carcinoma by age, W-county.

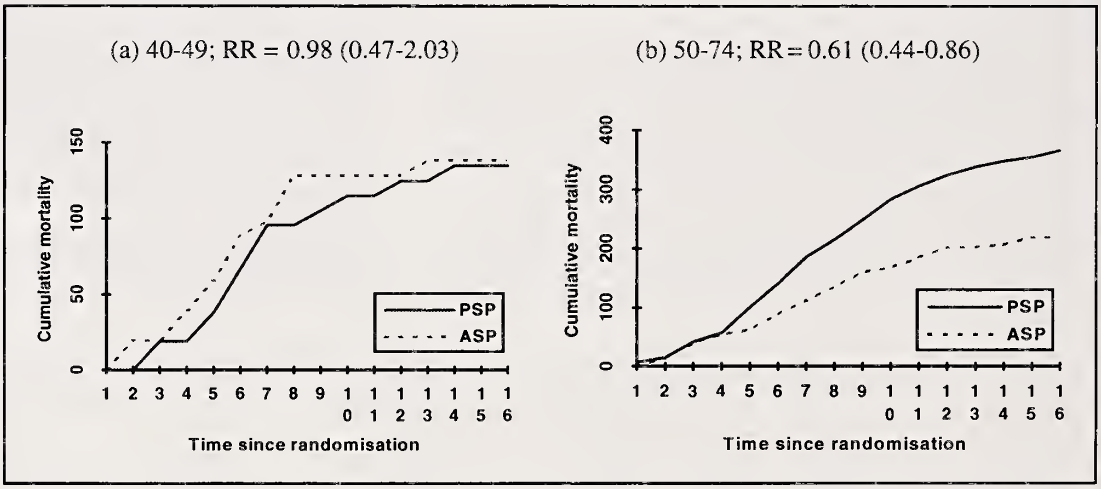

Fig. 4. Cumulative mortality from breast cancer for ductal-grade 3 carcinoma by age, E-county.

40–49 is likely to be more crucial than for women aged 50 and over (*15*). Using Markov-chain models based on tumor size, node status, and malignancy grade, we have recently demonstrated that when changing the screening interval from three years to one year, the proportion of tumors which are already advanced in their development (tumors of size 2 cm or more, node positive, and grade 3) may be reduced from 17% to 5% in women aged 40–49 but only from 9% to 4% and from 6% to 3% in women aged 50–59 and 60–69 respectively (*12*).

The good correlation between relative mortality (ASP versus PSP) and relative incidence of Stage II and worse tumors, as shown in Figs. 1 and 2, suggests that the relative incidence of tumors of Stage II+ is a good predictor of the subsequent effect on mortality. This is in accordance with previous findings (*5,6,7,8*). The relative mortality predicted from tumor size, node status, and malignancy grade has also been shown to agree well with observed relative mortality (*13*).

Our results point out the particular value of malignancy grade

in predicting how soon after initiating screening one can expect to see a mortality benefit. If indeed the screening program reduces the incidence of grade 3 ductal cancers and/or reduces the tumor size and frequency of nodal spread of the poorly differentiated ductal cancers, one could be confident that breast cancer mortality will be decreased and that an early benefit will be demonstrable. At the other extreme, early detection of DCIS cases and tubular, mucinous, and grade 1 ductal cancers will have little demonstrable effect on mortality within 10–15 years.

In conclusion, the results here suggest that the smaller and delayed benefit of two-year breast cancer screening in women aged under 50 years is mostly due to a small reduction in mortality from grade 3 ductal cancers. Progression of grade 3 carcinomas seems to be more rapid and dedifferentiation of low-grade cancers more frequent in younger women. This makes early detection more difficult, especially with a two-year screening interval. Accordingly, a shorter interscreening interval is required to detect these rapidly growing cancers at an earlier stage in their natural history.

References

(*1*) Tabar L, Dean PB. Breast cancer: a progressive, heterogeneous disease requiring multidisciplinary diagnosis and treatment. J Oncol Management 1994;Nov/Dec:12–3.

(*2*) Hellman S. Natural history of small breast cancers: Karnofsky Memorial Lecture. J Clin Oncol 1994;12:2229–34.

(*3*) Tubiana M, Koscielny S. Natural history of breast cancer: recent data and clinical implications. Breast Cancer Res Treat 1991;18:125–40.

(*4*) Tabar L, Fagerberg G, Day NE, Duffy SW, Kitchin RM. Breast canc treatment and natural history: new insights from the results of screening Lancet 1992;339:412–4.

(*5*) Shapiro S, Venet W, Strax P, et al. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69:349–55.

(*6*) Tabar L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Lancet 1985;I: 829–32.

(*7*) Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. Br Med J 1988;297:943–8.

(*8*) Frisell J, Eklund G, Hellstrom L, et al. Randomized study of mammography screening—preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18:49–56.

(*9*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Screening for breast cancer in women aged under 50: mode of detection, incidence and histology. J Med Screening 1995;2:94–8.

(*10*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age: new results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.

(*11*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Tumor development, histology and grade of breast cancers: prognosis and progression. Int J Cancer 1996;66:413–9.

(*12*) Chen HH, Duffy SW, Tabar L, Day NE. Markov chain models for progression of breast cancer: I. Tumor attributes and the preclinical screen-detectable phase. Submitted to Amer J Epidemiol.

(*13*) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer. In press.

(*14*) Breslow NE, Day NE. Statistical Methods in Cancer Research. Vol II. The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer, 1987.

(*15*) Tabar L, Fagerberg G, Day NE, Holmberg L. What is the optimum interval between mammographic screening examinations? An analysis based on the latest results of the Swedish Two-County breast cancer screening trial. Br J Cancer 1987;55:547–51.

The Stockholm Mammographic Screening Trial: Risks and Benefits in Age Group 40–49 Years

*Jan Frisell, Elisabet Lidbrink**

This article presents updated data on breast cancer mortality for women under age 50 from the Stockholm Mammographic Screening Trial, as well as a review of some side effects associated with screening in this age group. Approximately 40,000 women aged 40–64 (14,842 aged 40–49 years) were randomized to a trial of breast cancer screening by single-view mammography alone; 20,000 women (7,103 aged 40–49) were randomized to a control group. In the 40–49 age group, 24 and 12 breast cancer deaths were found in the study and control groups, respectively, after 11.4 years of follow-up. The relative risk of breast cancer death in screened to nonscreened women was 1.08 (95% confidence interval: 0.54–2.17). The rates of benign surgical biopsies, false positives, and follow-up costs were higher among women under age 50. Large overview studies are needed, however, to determine whether mammography screening consistently reduces mortality in women 40–49 years of age. Side effects such as costs and public health aspects of mammography screening in this age group also warrant further study. [Monogr Natl Cancer Inst 1997;22:49–51]

Results from several randomized mammography screening trials have shown that a mammographic screening program can reduce breast cancer mortality, at least for women above 50 years of age (*1,2,3,4*). For some years, however, there has been a perceived need for more information on the effect of screening in women aged 40–49 years. A Swedish overview of randomized mammographic screening trials found a relative mortality of 0.77 (95% confidence interval [CI]: 0.59–1.01) associated with screening, and meta-analysis combining all randomized trials gave a relative mortality of 0.85 (0.71–1.01) (*5*). It is likely that mammographic screening of women aged 40–49 can reduce subsequent mortality from breast cancer, but further work and analysis are needed before universal recommendations can be made for this age group, especially when one considers the side effects of mammography screening, such as false positives, follow-up costs, and benign biopsy rates. The aim of this report, therefore, is to present updated data on breast cancer mortality among women aged 40–49 years from the Stockholm Mammographic Screening Trial and also to examine some of the side effects associated with mammography screening in this age group.

Methods

Subjects and Screening

The Stockholm mammographic screening trial started in March 1981, when 40,318 women aged 40–64 years were randomized to a trial of breast cancer screening by single-view mammography alone versus no intervention in a control group of 20,000 women. The study was designed to have approximately twice as many subjects in the study group as the control group. At randomization, 14,842 women in the study group were aged 40–49 years compared to 7,103 in the control group. Two screenings rounds were performed, and the first and second screening intervals were 28 and 24 months respectively. Attendance was 81% in the first two screening rounds and equal in all age groups. The recall rates for complete mammography in age group 40–49 years were 5.1 and 4.0 in the first and second screening rounds respectively, and the recall rates for clinical examination, fine-needle biopsy, and complementary x-rays were 1.3 and 1.0 for the two rounds. During 1986, the control group was invited once to screening and the study was ended. The Stockholm mammographic screening trial is presented in detail in several reports (*6,7*).

Mammography

Mammography was performed with a CGR Mammograph (Senograph 500T). A single-view mammogram was obtained in oblique projection. If malignancy was suspected, the woman was recalled for a conventional three-view mammogram. Kodak NMB film was used with Kodak mammography cassettes and Kodak Min-R intensification screens. The film was exposed at 28 kv and developed at 34 °C for 2.5 minutes.

Statistical Methods

The mortality rates in the groups reflect the ratio between the number of deaths from breast cancer and the number of person-years. A log rank test was used to determine statistical significance, and the cut-off value was 0.05. All reported *P*-values are from two-sided tests. The significance analysis and the confidence intervals are based on the reasonable assumptions that the observed number of deaths are Poisson distributed and that the uncertainty in the number of person-years can be neglected. The relative risk is obtained as the mortality ratio between the study population and control population.

Assessment at the End of the Trial and Follow-Up

The endpoint in the trial was breast cancer deaths, which was defined as death with breast cancer present at death (locore-

**Affiliations of authors:* J. Frisell, Department of Surgery, Stockholm Söder Hospital, Stockholm, Sweden; E. Lidbrink, Department of Oncology, Stockholm Söder Hospital, Stockholm, Sweden.

Correspondence to: Jan Frisell, Department of Surgery, Stockholm Söder Hospital, S-118 83, Stockholm, Sweden.

gional or distant disease) (*4,7*). The "evaluation method" was used to calculate breast cancer deaths (*3*). This means that breast cancer deaths in the study and control groups among women who developed breast cancer after 1986 have not been included.

Results

Mortality Results

Figure 1 shows the cumulative number of deaths from breast cancer in relation to years after randomization for the 40–49 age group. In this age group, 118 and 59 breast cancers were diagnosed in the study and control groups respectively. After a mean follow-up of 11.4 years, there were 24 breast cancer deaths in the study group and 12 in the control group, with 173,866 and 87,826 person-years in the study and control groups respectively. The relative risk of breast cancer death was 1.08 (95% CI: 0.54–2.17). No mortality reduction was seen in this age group, in contrast to a significant mortality reduction among women over 50 years. The breakpoint for benefit in this study seemed to be at 50 years, but this tendency is uncertain because of the low statistical power in the analyses of small subgroups.

False Positives and Costs

The recall rate for clinical examination, fine needle biopsy, and additional x-rays after a complete mammography was 0.8% for all subjects and 1.0% in 40–49 age group. The number of false positives in relation to the number of cancers found in each age group was higher in the 40–49 age group compared to women over 50 years (Table 1). With only two screening rounds, the proportion of false positive cases was 242/100,000 women-years in women over 50 years compared to 355/100,000 women-years in women below 50 years. The rate of benign surgical biopsies in the incidence screening (second round) in age group 40–49 years was 49/100,000 women-years compared to 21/100,000 women-years among women over 50 years. One out of 2.5 surgical biopsies was benign in age group 40–49 years compared to one out of seven for women over 50 years. Forty-one percent and 56% of the follow-up costs of the false positives in the first and second screening rounds, respectively, resulted from examinations of women aged 40–49 (*8,9*).

Interval Cancers

Breast cancers diagnosed between two screening-rounds in the study group are called interval cancers. The incidence of interval cancers in the Stockholm study was 1.8 and 2.0 breast cancers/1,000 women/24 months in the first and second intervals respectively. There was a significantly larger number of younger women aged 40–49 years diagnosed with interval cancers ($P<0.05$). Among women aged 40–64 years, there was significantly better survival among the breast cancers diagnosed between two screening examinations compared to the control cancers ($P<0.01$) (*10*). This better survival, however, was confined to women over 50 years of age; in the 40–49 age group, survival was equal to the control group. The mortality of younger women in the study group was dominant (45.8%) among the interval cancers compared to 30% among screening-detected cancers and 30.4% among breast cancers diagnosed in the non-attenders group.

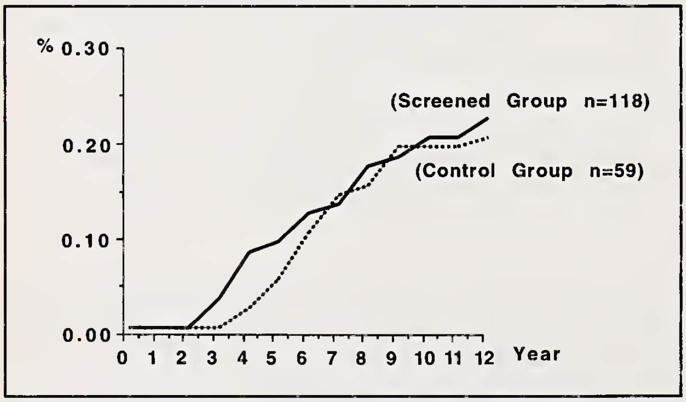

Fig. 1. Cumulative percent of deaths from breast cancers diagnosed in the study (n = 118) and control groups (n = 59), aged 40–49 at entry.

Discussion

In age group 40–49 years in the Stockholm trial, no mortality reduction was seen after 7.4 years and again after 11.4 years of follow-up (*4,7*). The breakpoint of benefit in this study seems to be 50 years, with a significant reduction in mortality among women over 50 years. This finding is preliminary, however, since the statistical power and the number of breast cancer deaths in age group 40–49 years was low. The results from the Stockholm trial were in line with the results from the two-county trial, but it is important to remember that none of the Swedish trials were designed to evaluate women aged 40–49 years as a separate group. Other trials, such as the Malmö and Gothenburg trials, have shown better survival for the younger women (*5*). Reasons for better survival in these trials were probably their shorter screening intervals and also the use of two-view screening in this age group. Longer intervals result in higher rates of interval cancer and higher mortality in women below 50 years.

The update report from the Swedish overview study presented at the Falun, Sweden, meeting in March 1996 has shown that it is possible to reduce mortality in the 40–49 age group, but the effect is lower compared to screening in women over 50 (*5*). These results from the Swedish overview study are in line with a recent meta-analysis of all population-based screening trials in the world (*5*). In order to reap any benefit in this age group, however, one must use high-quality, two-view mammography, 12- to 18-month screening intervals, and have high subject compliance.

The proportion of false positives was 47% higher in women under 50 than in older women, and the proportion of benign surgical biopsies was also higher (49/100,000 versus 21/100,000 women-years) in the younger age group. Needless to say, false positives foment patient anxiety and generate further costs. In the Stockholm study, the follow-up costs from false positive cases were not negligible, and in the second screening round, 56% of these costs belonged to the 40–49 age group (*8*).

Even if the Swedish overview study has shown a possible benefit of mammography screening in the 40–49 age group, further studies are needed to analyze side effects of mammography screening, such as false positives, costs, nonattendance, and mortality from interval cancers. Screening programs must achieve a balance between a possible mortality benefit and these potential risks before recommendations for mammography screening can be made to all women aged 40–49 years.

Table 1. Ratio of benign to malignant in detected breast lesions from the first and second rounds in relation to different age groups

Age (years)	First screening round			Second screening round		
	Benign	Malignant	Ratio	Benign	Malignant	Ratio
40–44	72	8	9.0	49	12	4.1
45–49	66	14	4.7	44	12	3.7
50–54	72	20	3.6	21	12	1.8
55–59	68	42	1.6	20	24	0.8
60–64	74	44	1.7	16	36	0.4
Total	352	128	2.8	150	96	1.6

References

(*1*) Shapiro S, Strax P, Venet L. Periodic breast cancer screening in reducing mortality from breast cancer. JAMA 1971;215:1777–85.

(*2*) Tabar L, Fagerberg G, Gad A, Baldetorp L, Holmberg L, Grontoft O, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985; 1:829–32.

(*3*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomized trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.

(*4*) Frisell J, Lidbrink E, Hellstrom L, Rutqvist LE. Follow-up after 11 years: update of mortality results in the Stockholm Mammographic Screening Trial. Breast Cancer Res Treat 1997. In press.

(*5*) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer 1996;68:693–9.

(*6*) Frisell J, Glas U, Hellstrom L, Somell A. Randomized mammographic screening for breast cancer in Stockholm. Design, first rounds results and comparisons. Breast Cancer Res Treat 1986;8:45–54.

(*7*) Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist LE, Somell A. Randomized study of mammographic screening—preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18: 49–56.

(*8*) Lidbrink E, Elfving J, Frisell J, Jonsson E. Neglected aspects of false positive findings of mammography in breast cancer screening: analysis of false positive cases from the Stockholm trial. BMJ 1996;312:273–6.

(*9*) Lidbrink E, Levi L, Pettersson J, et al. Single-view screening mammography: psychological, endocrine and immunological effects of recalling for a complete three-view examination. Eur J Cancer 1995;31A:932–3.

(*10*) Frisell J, von Rosen A, Wiege M, Nilsson B, Goldman S. Interval cancer and survival in a randomized breast cancer screening trial in Stockholm. Breast Cancer Res Treat 1992;24:11–6.

The Gothenburg Breast Cancer Screening Trial: Preliminary Results on Breast Cancer Mortality for Women Aged 39–49

*Nils Bjurstam, Lena Björneld, Stephen W. Duffy, Teresa C. Smith, Erling Cahlin, Olof Erikson, Halvard Lingaas, Jan Mattsson, Stellan Persson, Carl-Magnus Rudenstam, Johan Säwe-Söderberg**

We carried out a randomized trial of invitation to screening mammography in the city of Gothenburg, Sweden, to estimate the effect of screening on breast cancer mortality in women under age 50 years. A total of 11,724 women aged 39–49 were randomized to the study group, which was invited to mammographic screening every 18 months; 14,217 women in the same age range were randomized to a control group, which was not invited to screening until the fifth screen of the study group. Breast cancers diagnosed in both groups between randomization and immediately after the first screen of the control group were followed up for death from breast cancer to the end of December 1994. There was a significant 44% reduction in mortality from breast cancer in the study group compared to the control group (relative risk [RR] = 0.56, *P* = 0.042, 95% confidence interval [CI]: 0.32–0.98). A conservative estimate based on removal of the cancers detected at the first screen of the control group gave an RR = 0.59 (*P* = 0.069, 95% CI: 0.33–1.05). The true answer is likely to lie between the two estimates. These data suggest that mammographic screening can reduce breast cancer mortality in women under age 50, particularly if high-quality mammography is used and a short interscreening interval is adhered to. [Monogr Natl Cancer Inst 1997;22: 53–55]

The effect of mammographic screening for breast cancer in women under age 50 years is an issue of controversy, partly due to the lesser effect on breast cancer mortality observed in this age group than in older women, and partly to the variation in estimated effects between randomized trials (*1–7*). Meta-analyses have not resolved the issue: even when several trials are combined there is still a relatively small number of breast cancer deaths in this age group, so that confidence intervals remain wide (*8–11*). There is evidence that interval cancer rates are higher in this age group (*12*), and that screening sensitivity is lower (*11*), probably due to a high prevalence of mammographically dense tissue in premenopausal women. These findings suggest that in women under age 50, screening has to be more frequent than in older women, and that measures must be taken to minimize the number of false negative screens—measures such as careful attention to mammographic quality, two-view mammography, and double reading (*11*).

Subjects and Methods

The subjects in this trial were the entire female population of the city of Gothenburg born between the years 1923 and 1944 inclusive. All women with a history of breast cancer prior to randomization were excluded. There were 51,611 women aged 39–59 at randomization. In this paper, we restrict analysis to the 25,941 women aged 39–49.

We planned to screen every 18 months. With the resources available, this dictated that the group invited to screening must number around 21,000 women. We therefore randomized to the study or control group in a ratio of 1 to 1.2 in the 39–49 age group and 1 to 1.6 in the 50–59 group. Randomization took place within each year of birth cohort successively. Thus, the 1923 cohort was randomized in December 1982 and the study group members invited to their first screen between December 1982 and February 1983. The last cohort to be randomized, women born in 1944, was randomized in April 1984 and the study group members received their first invitation in May, 1984. The randomization was by cluster, based on day of birth in the 1923–1935 cohorts, and by individual for the 1936–1944 cohorts, as the computer software for screening invitation was amended during the period of the trial to enable individual randomization. In the 39–49 age range, the final sample comprised 11,724 in the study group and 14,217 in the control group. The mean ages in the study and control groups were 43.9 years and 43.8 years respectively.

The study group members were invited to screening every 18 months. The control group members received a single screen immediately following the fifth screen in the study group. The cancers diagnosed from the time of randomization up to immediately after the first screen of the control group (which was completed on average around seven years after randomization) were then followed up for breast cancer mortality.

The screening modality was mammography. Two-view mam-

**Affiliations of authors:* N. Bjurstam, L. Björneld, H. Lingaas, Department of Diagnostic Radiology, Sahlgrens University Hospital, Gothenburg, Sweden; S. W. Duffy, T. C. Smith, MRC Biostatistics Unit, Cambridge, UK; E. Cahlin, J. Mattsson, C. M. Rudenstam, Department of Surgery, Sahlgrens University Hospital, Gothenburg, Sweden; O. Erikson, S. Persson, J. Säwe-Söderberg (now deceased), Department of Cyto-Pathology, Sahlgrens University Hospital, Gothenburg, Sweden.

Correspondence to: Dr. Nils Bjurstam, Department of Diagnostic Radiology, Sahlgrens University Hospital, S-41345, Gothenburg, Sweden.

mography was used at the first screen, and single view at later screens, unless the density of the breast at the first screen indicated that single-view mammography would be inadequate. Screening took place in a stationary unit with specially trained radiology nurses. Mammography was performed using a unit with CGR Senograph 500 T with moving grid. We used the Kodak Min R imaging system, with extended film processing (three minutes). Films were single read at the first three screening rounds and double read thereafter, and those recalled were subject first to supplementary mammography, and, if necessary, to physical examination by a surgeon and to fine needle aspiration cytology.

The primary outcome was mortality from breast cancers diagnosed during the period of the trial, as defined above. Mortality data were available up to December 31, 1994. Breast cancer deaths were identified from the Swedish cause of death register, which was shown in the overview of Swedish breast screening trials to be reliable (*13*). Mortality was compared between the study and control groups using Poisson regression (*14*).

Results

Table 1 shows attendance rates and cancers diagnosed at each screen. Attendance was between 75% and 85% in the study group and was 66% at the first screen of the control group. The cancer detection rate at the first screen of the control group was higher than that for the study group, as the women in the eontrol group were on average six years older at their first screen.

During the screening period of the trial, there were 144 breast cancers diagnosed in the study group and 195 breast cancers in the control group. There were 18 deaths and 138,402 person-years to the end of 1994 in the study group, and 39 deaths and 168,025 person-years in the control group. A signifieant reduction in mortality was observed in the study group ($P = 0.042$), with a relative risk (RR) of 0.56. Figure 1 shows cumulative mortality over time in the study and control groups. The mortality of the two groups began to separate between six and eight years after randomization, and the gap continued to widen thereafter.

Table 2 gives the cancer incidence during the screening phase of the trial. There was a 10% lower incidence in the study group. This difference was not significant (but see Discussion below).

Discussion

The results above are consistent with previous findings of reduced breast cancer mortality from screening women under age 50 years (*11*). The present results also suggest that with high-quality mammography and a short screening interval, the benefit can be substantial. This is the first internal analysis of mortality in the Gothenburg trial, and further follow-up is necessary to ensure that the mortality benefit is maintained. Since a number of screens were performed after age 50, further analyses are required to determine the magnitude of the benefit from screening with respect to cancers diagnosed before age 50.

Although the difference in breast cancer incidence between the study and control groups is not significant (relative incidence = 0.90; 95% confidence interval [CI]: 0.72–1.13), it is advisable to consider the possibility of bias. For example, the higher incidence in the control group may be due to the fact that the first screen of the control group ended on average a few months later than the fifth screen of the study group. Indeed, it is at this first screen of the control group that the excess incidence in the control group occurs, as shown in Table 1. Because the closure of the screening phase of the trial occurred at the same point in time for both the study and control groups, there is more opportunity for lead time cancers to be diagnosed and therefore followed up in the control group than in the study group (due to the later screen).

To obtain a more conservative estimate, we excluded all cancers in the control group diagnosed after the start of screening the control group, and therefore the five breast eancer deaths among these. This left 151 breast cancers and 34 breast cancer deaths in the control group. However, without an adjustment to the person-years, this exclusion would have biased the results in the opposite direction, with a considerable deficit of cancers in the control group. We therefore made the following adjustment to the person-years: since we had additional cancers in the study group due to lead time from the final screen, and no such cancers in the control group, we added the corresponding person-years to that of the study group. Using the method of Paci and Duffy (*15*), we estimated the expected lead time as 2.21 years and the sensitivity as 0.87. The additional number of person-years equaled the number screened at the final screen of the study group times the sensitivity (0.87) times 2.21– 0.81 (the lead time minus the time from screening the study group to closing the recruitment period in years). Thus, we added $1.4 \times 8{,}675 \times 0.87 = 10{,}566$ to the person-years of the study group. We also subtracted from the person-years in the control group 0.19 (the average time in years taken to screen each year of birth cohort in the control group) times the number in the control group at the time of the invitation, 13,947. This gave a total of 91,907 person-years in the study group during the caneer recruitment period and 96,098 in the eontrol group. The relative incidence was now 1.00. Performing the same adjustment to the person-years of total follow-up and recalculating the breast cancer mortality, we arrived at 18 deaths and 148,968 person-years in the study group, 34 deaths and 165,375 person-years in the control group, and an RR of 0.59 ($P = 0.069$; 95% CI: 0.33–1.05). This is likely to be conservative, as it involves removing 23% of the cancers in the control group but adjusting the person-years for mortality by only 8% in the study group and 2% in the control group. The true RR may lie between the 0.59 calculated here and the 0.56 given above.

The attendance rates in the study group were between 75% and 85%, in line with other Swedish programs (*9*). In a survey

Table 1. Attendance rates and diagnostic work-up by screening round

Screening round	Invited	Attended (%)	Cancers (%)
1 (Study)	11,720*	9,921 (85)	17 (0.17)
2 (Study)	11,679	9,157 (78)	10 (0.11)
3 (Study)	11,624	9,150 (79)	15 (0.16)
4 (Study)	11,571	8,914 (77)	21 (0.23)
5 (Study)	11,519	8,675 (75)	20 (0.23)
1 (Control)	13,947*	9,167 (66)	40 (0.44)

*Numbers are smaller than total cohort because of losses between randomization and first screen.

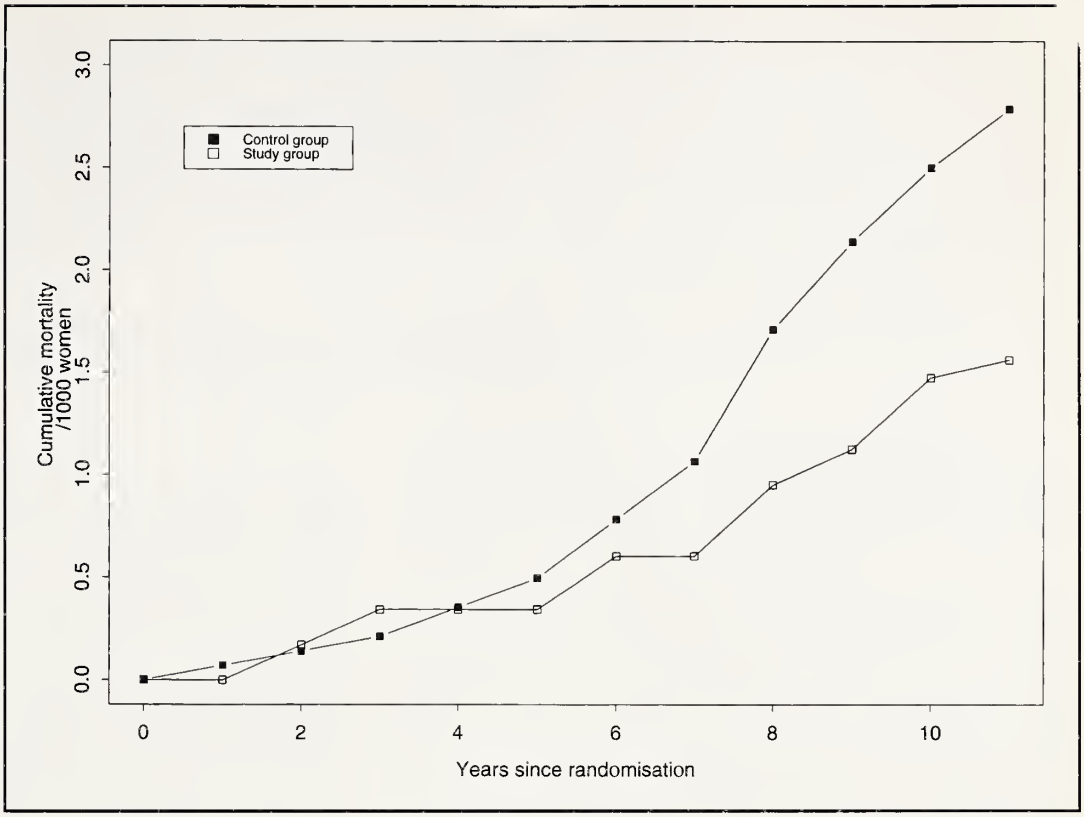

Fig. 1. Cumulative breast cancer mortality in women aged 39–49 at randomization.

Table 2. Incidence of breast cancer in study and control groups during the screening phase of the trial

Group	Breast cancers	Person-years (in screening phase)	Relative incidence (95% CI)
Control	195	98,748	1.00 (–)
Study	144	81,341	0.90 (0.72–1.12)

of 1,641 controls in this age group, 19% reported having had a mammogram in the last two years. Thus, there may have been some "voluntary" screening in Gothenburg before and during the trial, and the mortality benefit observed in this trial is likely to be a result of enrollment in an organized program with a strict 18-month interscreening interval and high-quality mammography. This is further supported by the fact that 33% of the deaths in the study group were from the nonattenders.

In conclusion, this trial adds to the evidence of a reduction in breast cancer mortality in women under age 50 invited for mammographic screening, and suggests that a substantial mortality benefit can result from a strict 18-month interval between screens.

References

(*1*) Shapiro S, Venet W, Strax P, Venet L. Periodic Screening for Breast Cancer: the Health Insurance Plan Project and its Sequelae 1963–86. Baltimore: Johns Hopkins University Press, 1988.

(*2*) Alexander FE, Anderson TJ, Brown HK, Forrest APM, Hepburn W, Kirkpatrick AE, et al. The Edinburgh randomised trial of breast cancer screening: results after 10 years of follow-up. Br J Cancer 1994;70:542–548.

(*3*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age: new results from the Swedish two-county trial. Cancer 1995;75:2507–2517.

(*4*) Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. Br Med J 1988;297:943–948.

(*5*) Frisell J, Eklund G, Hellstrom L, Lindbrink E, Rutqvist LE, Somell A. Randomized study of mammographic screening—preliminary report on mortality in the Stockholm Trial. Breast Cancer Res Treat 1991;18:49–56.

(*6*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. Can Med Assoc J 1992;147:1459–1476.

(*7*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. Can Med Assoc J 1992;147:1477–1488.

(*8*) Smart CR, Hendrick RE, Rutledge JH, Smith RA. Benefit of mammography screening in women aged 40 to 49 years: current evidence from randomized trials. Cancer 1995;75:1619–1626.

(*9*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised controlled trials. Lancet 1993;341:973–978.

(*10*) Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50: a quality assessment and meta-analysis. Med J Austral 1995;162:625–629.

(*11*) Organizing Committee and Collaborators, Falun meeting. Breast cancer screening with mammography in women aged 40–49 years. Int J Cancer 1996;68:693–699.

(*12*) Tabar L, Fagerberg G, Day NE, Holmberg L. What is the optimum interval between mammographic screening examinations? An analysis based on the latest results of the Swedish two-county breast cancer screening trial. Br J Cancer 1987;55:547–551.

(*13*) Nystrom L, Larsson LG, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, et al. Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials: a comparison between official statistics and validation by an endpoint committee. Acta Oncologica 1995;34:145–152.

(*14*) Breslow NE, Day NE. Statistical Methods in Cancer Research II. The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer, 1987.

(*15*) Paci E, Duffy SW. Modeling the analysis of breast cancer screening programmes: sensitivity, lead time and predictive value in the Florence District Programme (1975–1986). Int J Epidemiol 1991;20:852–858.

n

Updated Overview of the Swedish Randomized Trials on Breast Cancer Screening With Mammography: Age Group 40–49 at Randomization

*Lars-Gunnar Larsson, Ingvar Andersson, Nils Bjurstam, Gunnar Fagerberg, Jan Frisell, László Tabár, Lennarth Nyström**

The purpose of this overview is to estimate more precisely the long-term effect of mammography screening by adding four more years of follow-up to women aged 40–49 years in the four Swedish trials on mammography screening. Data from the four trials were merged and linked to the Swedish Cancer and Cause of Death Register for 1958–1993 and 1951–1993 respectively to identify date of breast cancer diagnosis and cause and date of death. The invited and control groups comprised 48,569 and 40,247 women respectively. At the December 1993 follow-up, 602 and 482 breast cancer cases were identified in the two groups respectively, of which 104 and 111 had breast cancer as the underlying cause of death. This corresponds to a relative risk (RR) of 0.77 (95% CI: 0.59–1.01) for the two groups. In the 40–44 age group at randomization, 94% of breast cancer patients in the study and 89% in the control group were diagnosed before the age of 50; however, among breast cancer deaths in this age group, only two in the invited and five in the control group died after age 50. At follow-up of women 40–44 years at randomization 208 women in the invited and 184 in the control group were reported to the Cancer registry with breast cancer. Out of these 195 (94%) and 163 (89%) respectively were reported before the age of 50. Further, the relative risk for the age group 40–44 years at randomization by age at follow-up was 1.11, 0.51 and 0.46 for the age groups 45–49, 50–54, and 55–59 at follow-up. This study shows a 23% reduction in the breast cancer mortality in women 40–49 years at randomization achieved from a median trial time of 7.0 years, a median follow-up time of 12.8 years, and a screening interval of 18–24 months. Almost all of the effect in the 40–44 year age group at randomization was due to screening before the age of 50. [Monogr Natl Cancer Inst 1997;22:57–61]

Four out of seven randomized controlled trials on mammography screening for breast cancer have been performed in Sweden. These trials—conducted at Malmö, Kopparberg/Östergötland (the "two-county trial"), Stockholm, and Gothenburg—contain about 60% of the subjects in such screening studies. The four trials were similarly designed. Each was population based, used mammography alone with one or two projections as a primary screening modality, and had a screening interval of 18–33 months. Three of the these trials suggest a reduction of the breast cancer mortality, but this reduction was statistically significant in only one of the trials (the two-county trial). Moreover, the efficacy among women aged 40–49 was uncertain.

To improve the precision in the estimates and to facilitate age stratification, we performed an overview (meta-analysis using individual patient data) and found at follow-up December 31, 1989 a 24% statistically significant overall reduction in breast cancer mortality among those invited to mammography screening. (*1*). The mortality reduction was similar irrespective of the endpoint used for evaluation, whether "breast cancer as underlying cause of death" or "breast cancer present at death." There was a consistent risk reduction associated with screening in all individual studies, although the point estimate of the relative risk (RR) for all ages varied nonsignificantly between 0.68 and 0.84. The largest reduction of breast cancer mortality, 29%, was observed among women aged 50 to 69 at randomization. There was a nonsignificant 13% reduction among women aged 40–49 at randomization. However, the cumulative breast cancer mortality curves seemed to diverge after about eight years, and, as both the Gothenburg and the Stockholm trials had a rather short follow-up time and contained more than half of women aged 40–49 years at invitation, we decided that a prolonged follow-up could increase the knowledge about the effect in this age group.

The aim of the present study, then, is to gain more precision in the RR estimates and to provide information on the long-term effect of breast cancer screening with mammography by adding four more years of follow-up. The analysis will focus on the age group 40–49 years at randomization.

**Affiliations of authors:* L.-G. Larsson (Chairman), Department of Oncology, Umeå University, Umeå, Sweden; I. Andersson (Trialist), Department of Diagnostic Radiology, Malmö Hospital, Malmö, Sweden; N. Bjurstam (Trialist), Department of Diagnostic Radiology, Sahlgrenska Hospital, Gothenburg, Sweden; G. Fagerberg (Trialist), Department of Diagnostic Radiology, University Hospital, Linköping, Östergötland County, Sweden; J. Frisell (Trialist), Department of Surgery, South Hospital, Stockholm, Sweden; L. Tabár (Trialist), Department of Mammography, Falun Hospital, Falun, Kopparberg County, Sweden; L. Nyström (Project Coordinator, Statistician), Department of Epidemiology and Public Health, Umeå University, Umeå, Sweden.

Correspondence to: Lennarth Nyström, Department of Epidemiology and Public Health, Umeå University, S-901 85 Umeå, Sweden.

See "Note" following "References."

Material and Methods

Study Design

The basic characteristics of the four randomized trials on mammography screening in Sweden have been extensively presented before, and a summary was presented in our first report from the overview (*1*). Initially, each screening center sent to the department of Epidemiology and Public Health in Umeå a magnetic tape containing the following information for each woman in the cohort: personal identification number, date of randomization, and date when the first round was completed for the control group. The cohorts were merged and linked 1) to the six Regional Cancer Registers to identify cases with breast cancer diagnosed between 1958 and 1993, and 2) to the Swedish Cause of Death Register at Statistics Sweden to identify women who died between 1951 and 1993 and the cause of death respectively. The Swedish Cancer and Cause of Death Registers have been shown to accurately record breast cancer data (*2*).

Exclusion Criteria

All analyses in the present study were based on exact age at randomization, despite the fact that most trials used—for practical reasons—year-of-birth cohorts. Thus, 5143 women aged 38–39 years at randomization were excluded (Kopparberg = 1148, Östergötland = 1296, Stockholm = 680, and Gothenburg = 2019), as we focused on the age group 40–49 years at randomization.

Women with breast cancer diagnosed before the date of randomization, according to the Swedish Cancer Register, were excluded from the cohorts (invited group (IG) = 272, control group (CG) = 256).

Determination of Cause of Death

In the original overview (*1*), cause of death was determined by an independent endpoint committee (EPC) consisting of four physicians who blindly reviewed medical records, autopsy protocols, cause of death certificates, and histopathology reports for all deceased breast cancer cases—that is, breast cancer (ICD = 174) patients who were reported to the Cancer Register after randomization and who died before follow-up. Later, the RR estimates according to the EPC were compared to the estimates according to the Cause of Death Register at Statistics Sweden for both models of analysis (see below) (*2*). The RRs determined by these methods were, for both models, very similar, but with a slight tendency towards higher values when Statistics Sweden was used. Since using Statistics Sweden to determine cause of death provides a conservative estimate of the effect of screening, we decided to use it in the present study.

Models for Analysis

Later in each trial, the control group was also invited to one screening before the trial terminated. Therefore, two different models were used for evaluation—the "follow-up" model and the "evaluation" model (*1*). The former model included all breast cancer deaths that occurred among women with a primary diagnosis after the date of randomization and before the common fixed study endpoint at December 31, 1993. The latter model ignored breast cancer deaths among women whose primary tumor, according to the Cancer Register, was diagnosed after completion of the first screening round of the control group. In the first follow-up, held until December 31, 1989, the "follow-up" and the "evaluation" models showed similar results. In the second follow up until December 31, 1993 only the "evaluation" model was used, since the duration from the completion of the first screening round of the control group to the date of follow-up had increased considerably, and the "follow-up" model thereby would result in biased estimates.

Statistical Methods

Statistical and epidemiological data analyses have been performed using the QUEST software program (*4*). RRs have been calculated using the density method, whereby the person-time experience of the cohort by time interval of follow-up is used to estimate the mortality rates in breast cancer. Weighted RRs and confidence intervals (CIs) have been calculated using Mantel-Haenszel procedures.

Results

Table 1 presents the number of women by age at randomization and by screening center. After exclusion of those who were reported to the Cancer Register with breast cancer before randomization, the invited and control groups comprised 48,569 and 40,247 women respectively. Of these, 1128 and 849, respectively, were reported to the Cancer Register with breast cancer before December 31, 1993, the end of the follow-up period. When excluding breast cancer cases diagnosed after the first screening round of the control group ("evaluation" model), 602 and 482 breast cancer cases remained in the overview material. Among these, 129 in the invited group and 128 in the control group died during the follow-up period. Breast cancer was the underlying cause of death in 104 women invited to screening and 111 women in the control group.

The time difference from the date of randomization until the end of first screening round of the control group varied from 4.4 years in Stockholm to 14.6 years in Malmö, resulting in a median trial time of 7.0 years (Table 2).

The median follow-up time—that is, the time from date of randomization until the end of follow-up (12/31/93)—was 12.8 years, varying from 9.9 in Gothenburg to 15.5 in Malmö (Table 2).

Table 3 shows the number of person-years of follow-up time in the invited group (616,264) and in the control group

Table 1. Number of women by age at randomization: invited (IG), control group (CG), and screening center

	Age at randomization					
	40–44		45–49		40–49	
Screening center	IG	CG	IG	CG	IG	CG
Malmö			3945	4017	3945	4017
Kopparberg	4595	2478	5055	2531	9650	5009
Östergötland	5157	5337	5062	5074	10240	10411
Stockholm	7517	4495	6668	3470	14185	7985
Gothenburg	5664	7106	5157	5995	10821	13101
Overview	22954	29416	25887	21087	48841	40503

Table 2. Median and lower-upper limit (L-UL) in years for trial time (from date of randomization until the first control group round was screened) and follow-up time (from date of randomization until date of follow-up (12/31/93) by screening center

Screening center	Trial time: Median	Trial time: L-UL	Follow-up time: Median	Follow-up time: L-UL
Malmö	14.6	13.9–15.2	15.5	15.3–16.8
Kopparberg	7.1	5.2–7.5	15.2	13.9–16.5
Östergötland	7.6	6.5–8.7	14.2	12.8–15.6
Stockholm	4.4	3.2–4.9	11.9	10.6–12.8
Gothenburg	7.0	6.6–7.5	9.9	9.7–10.3
Overview	7.0	3.2–15.2	12.8	9.7–16.8

Table 3. Number of 1000 person-years and number of cases with breast cancer as the underlying cause of death according to Statistics Sweden

Screening center	No. of 1000 person-years: IG	No. of 1000 person-years: CG	No. of deaths: IG	No. of deaths: CG	RR	95% CI
Malmö	61	62	15	23	0.67	0.35–1.27
Kopparberg	144	75	23	18	0.67	0.37–1.22
Östergötland	143	147	27	27	1.02	0.59–1.77
Stockholm	162	94	23	10	1.34	0.64–2.80
Gothenburg	106	129	16	33	0.59	0.33–1.06
Overview:						
40–44	283	235	39	44	0.74	0.48–1.14
45–49	334	272	65	67	0.79	0.56–1.11
40–49	616	506	104	111	0.77	0.59–1.01

(506,358). During follow-up, 104 and 111 breast cancer deaths respectively occurred. This corresponds to a mortality reduction of 23% (RR = 0.77; 95% CI: 0.59–1.01). The effect was similar in the age cohorts 40–44 and 45–49 years at randomization: 26% and 21% respectively.

Figure 1 demonstrates the cumulative breast cancer mortality curves per 100,000 person-years by time since randomization. For the age group 40–49 years at randomization, the curves start to diverge after about six years and continue to diverge at 15 years of follow-up. The effect in the age group 40–44 and 45–49 at randomization was almost identical. Notice that the latest trial (Gothenburg) only contributes to the first 10 years (median follow-up time is 9.9 years).

An important question is whether the impact of screening on breast cancer mortality among women aged 40–49 at randomization originates from cases diagnosed before or after 50 years of age. Table 4 shows that 54% (326/602) of the invited group were younger than 50 years at the time of diagnosis, whereas 50% (240/482) of the control group were diagnosed before the age of 50. The corresponding figures for the breast cancer deaths were 60% and 53% in the invited and control groups respectively. However, for the age group 40–44 years at invitation, 94% of all cases in the invited group and 89% in the control group were younger than 50 years when they were reported to the Cancer Register with breast cancer, and only two in the invited group and five in the control group were reported to the Cause of Death Register with breast cancer as the underlying cause of death after the age of 50.

Another way to address this question is to calculate the RR by age at follow-up (Table 5). For women 40–44 years at randomization and 45–49, 50–54 and 55–59 years at follow-up the RR was 1.11, 0.51 and 0.465 respectively. Since the median trial time was 7.0 years, the mortality decrease described above could not have originated from diagnosis after the age of 50. In the age group 45–49 at randomization, the effect seems to appear earlier but at a lower level than for the 40–44 year age cohort.

Discussion

The mortality reduction shown in the present overview is close to statistically significant. There are, however, differences between the studies in terms of the number of screening rounds, the screening interval, and initiation date—and consequently

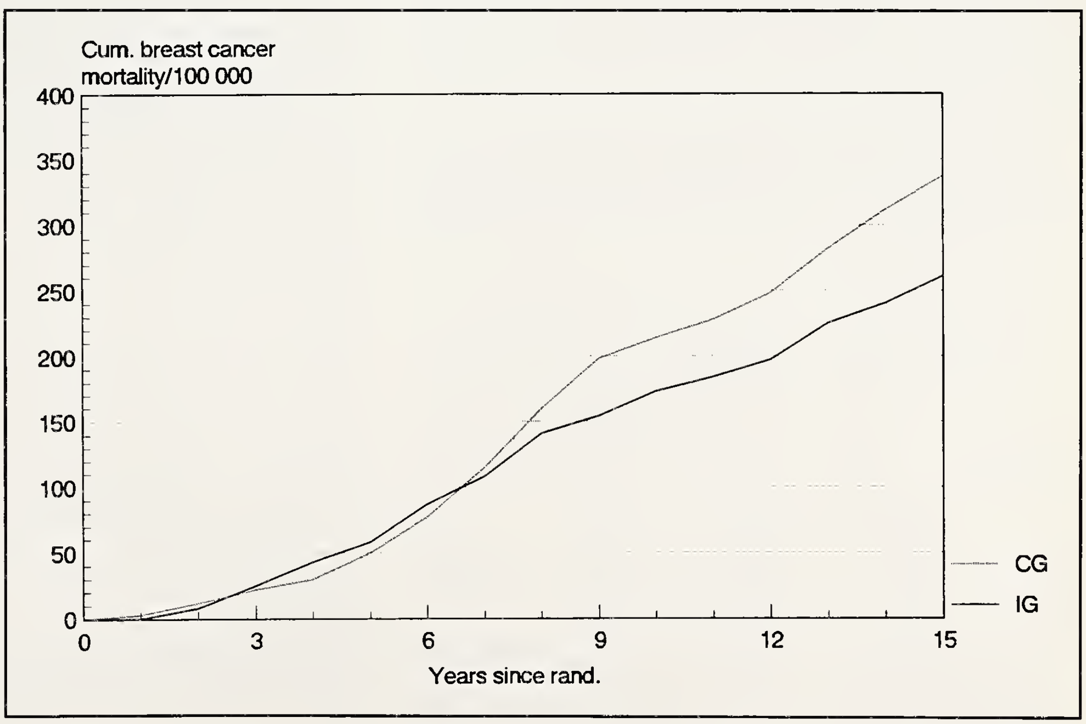

Fig. 1. Cumulative number of breast cancer deaths (BCDs)/ 100,000 women 40–49 years at randomization in the invited (IG) and control group (CG) by year since randomization.

Table 4. Number of breast cancer (BC) cases and breast cancer deaths by age at diagnosis in invited group (IG) and control group (CG) for women 40–49 years at randomization

Age at randomization	Group	Age at breast cancer diagnosis ≤49	≥50	No. BC diagnosis*	Total
Breast cancer cases:					
40–44	IG	195	13	—	208
	CG	163	21	—	184
45–49	IG	131	263	—	394
	CG	77	221	—	298
40–49	IG	326	276	—	602
	CG	240	242	—	482
Breast cancer deaths:					
40–44	IG	35	2	2	39
	CG	37	5	2	44
45–49	IG	25	38	2	65
	CG	19	45	3	67
40–49	IG	60	40	4	104
	CG	56	50	5	111

*According to the Cancer Registry.

some differences in screening modalities. The Gothenburg trial, for instance, used a short screening interval (18 months), completed five screening rounds, was initiated last, and had the shortest follow-up; therefore it used a grid technique from the beginning. On the contrary, the Stockholm trial completed only two rounds with non-grid technique and had a screening interval of about two years. Thus, these studies—especially the Gothenburg trial—may show a further reduction of the breast cancer mortality with longer follow-up.

Extending the follow-up until 1993 raises the question of whether the screening effect is statistically significant in the age group 40–49 years at randomization when the younger trials in Gothenburg and Stockholm have been followed for 15 years. However, following a cohort aged 40–49 years at randomization for 15 years can obscure the impact of any screening done before the age of 50. In their analysis of the 40–49 year age group in the two-county trial, Tabár et al. (*7*) found that "[f]or cancers diagnosed before age 50 years, the relative mortality, adjusted for age at randomization and county, was 0.85. For cancers diagnosed after age 50 years, the relative mortality was 0.95."

Table 5 shows that almost all of the mortality benefit observed in women aged 40–44 at randomization is due to cancers diagnosed before age 50. For women aged 45–49 at randomization, some of the benefit appears to accrue from cancers diagnosed after age 50. Paradoxically, the more efficient the screening before age 50, the more tumors will be diagnosed before age 50 in the invited group (and hence the fewer breast cancer deaths in the invited group after age 50). Thus, effective screening before age 50 results in a reduced mortality in the invited group from tumors diagnosed after age 50.

Table 5. Incidence of breast cancer deaths in invited group (IG) and control group (CG) and relative risks (RRs) by age at randomization and age at follow-up

Age at randomization		Age at follow-up 40–44	45–49	50–54	55–59	60–64	40–64
40–44	IG	0	2.20	1.26	1.28		1.38
	CG	0.58	1.98	2.46	2.77		1.87
	RR	0	1.11	0.51	0.46		0.74
45–49	IG		0.96	2.35	1.82	2.55	1.95
	CG		0.60	2.79	2.82	3.96	2.54
	RR		1.60	0.84	0.65	0.64	0.77
40–49	IG	0	1.76	1.91	1.92	1.72	1.69
	CG	0.58	1.51	2.66	2.82	3.94	2.83
	RR	0	1.17	0.72	0.61	0.64	0.76

Quantifying screening effect relative to patient age is, however, a complex issue confounded by the lead time in the invited group. As Table 4 shows, in the age group 40–49 at randomization, 54% (326/602) of tumors in the invited group and 50% (240/482) in the control group were diagnosed before age 50. Thus, the number of cancers in the invited group that would have been detected after age 50 in the absence of screening can be estimated as 602 × 242/482 = 302, 9.5% more than the 276 observed. If we apply this to the breast cancer deaths, then there are four less deaths (40 × 0.095) prevented in cancers diagnosed after age 50 than would appear from Table 5. These deaths have not been prevented by screening after age 50, but have been moved from diagnosis after age 50 to diagnosis before age 50 by lead time. This gives 56 deaths from the cancers diagnosed before age 50. If we use the person-years from Table 3, we obtain a 0.82 relative mortality, approximately adjusted from lead time, for cancers diagnosed before age 50, and 0.72 for cancers diagnosed after age 50. The 0.82 agrees exactly with the benefit of screening in women aged 40–49 every two years as estimated from modeled effects of screening on tumor size and node status (*8*), and the 0.72 is in line with the observed effect in women aged 50–59 at randomization (*6*).

It has been shown (*5*) that the cause of death pattern in the invited group in these trials is very similar to that in the control group, except for in the case of breast cancer. This demonstrates that the groups are comparable. In addition, the total mortality in the control group, including the breast cancer mortality, is almost identical to that of Swedish women in general. The same is true for the invited group, with the exception of breast cancer. This confirms that the trial cohorts are representative of Swedish women, indicating that the quantitative results from these trials may safely be generalized to the Swedish population.

An alternative approach to estimating the effect of mammography screening was recently applied on the follow-up data prior to 1989 (*6*). By using official national cause of death statistics according to Statistics Sweden as a reference to estimate the breast cancer mortality in the breast cancer subcohorts, we obtained estimates very similar to the traditional comparison of the breast cancer mortality in the invited and control groups. This analysis further strengthens the previous report (*1*) of a beneficial effect of mammography screening.

Chen et al. (*9*) have calculated the number of breast cancer deaths that could be prevented per 10,000 women invited to screening. According to the WE-study, among women 40–44 and 45–49 years at invitation, 5.7 and 6.6 breast cancer deaths per 10,000 invited could be avoided as compared with 19 and 22 among women 50–64 and 65–74 years at invitation (Table 6) (*9*). The lower figure in younger women is due to a lower breast cancer incidence.

To conclude, this follow-up of the four randomized controlled trials on mammography screening for breast cancer in Sweden, in which women were screened for 7.0 years and followed up

Table 6. Avoided number of breast cancer deaths (BCDs)/10,000 women 40–49 years at randomization, with a median trial time of 7.0 years, a median follow-up time of 12.8 years, and 18- to 24-month screening interval

Age at randomization	No. of women		No. of BCDs		Expected no. in CG	Avoided no. of BCD
	IG	CG	IG	CG		
40–44	22954	19416	39	44	52*	5.7**
45–49	25887	21087	65	67	82	6.6
40–49	48841	40503	104	111	134	6.1

*22954(44/19416) = 52.02.
**(52-39)/22954 = 5.66/10,000.

for 12.8 years, indicates a possible—although not statistically significant—effect in women 40–49 years at randomization, and almost all of the effect is due to screening before the age of 50 years.

References

(1) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised studies. Lancet 1993;341:973–8.

(2) Garne JP, Aspegren K, Moller T. Validity of breast cancer registration from one hospital into the Swedish National Cancer Registry 1971–1991. Acta Oncol 1995;34:153–6.

(3) Nystrom L, Larsson LG, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, et al. Determination of cause of death among breast cancer cases in the Swedish randomised mammography screening trials. A comparison between official statistics and validation by an endpoint committee. Acta Oncol 1995; 34:145–32.

(4) Gustafsson L. QUEST. A program system for statistical and epidemiological data analysis. Umeå University, Umeå, Sweden. January 1990.

(5) Nystrom L, Larsson LG, Wall S, Rutqvist LE, Andersson I, Bjurstam N, et al. The overview of the Swedish randomised mammography trials. The total mortality pattern and the representativity of the study cohorts. J Med Screen 1996;3;85–87.

(6) Larsson LG, Nystrom L, Wall S, Rutqvist LE, Andersson I, Bjurstam N, et al. The Swedish randomised mammography screening trials. An analysis of their effect on the breast cancer related excess mortality. J Med Screen 1996;3:129–32.

(7) Tabar L, Duffy SW, Chen HH. Re: Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials. J Natl Cancer Inst 1996;88:52–3.

(8) Breast cancer screening with mammography in women aged 40–49 years. Report of the Organizing Committee and Collaborators, Falun Meeting, Falun, Sweden (1966 March 21–22). Int J Cancer 1996;68:693–9.

(9) Chen H-H, Tábár L, Fagerberg G, Duffy S. Effect of breast cancer screening after age 65. J Med Screen 1995;2:10–4.

Note

Supported by the Swedish Cancer Society.

Reduced Breast Cancer Mortality in Women Under Age 50: Updated Results From the Malmö Mammographic Screening Program

*Ingvar Andersson, Lars Janzon**

This article provides additional follow-up data of two cohorts from the Malmö Mammographic Screening Trial (MMST). The first cohort, MMST I, contained 7,984 women under age 50 at entry into MMST who were born between 1927 and 1932. Half were assigned to a control group and were not invited for examination until four years after the code was broken in the MMST in 1988. The second cohort, MMST II, contained 17,786 women born between 1933 and 1945. Fifty four percent of these women were randomly invited to screening between 1978 and 1990. The remaining 46%—the control group—was invited to screening between 1991 and 1994. Nine screening rounds were completed in MMST I, and a mean of five rounds were completed in MMST II; the screening interval ranged from 18 to 24 months. The effect of screening on breast cancer mortality was assessed by pooling the two cohorts. At the end of follow-up—December 1993 for MMST I and December 1995 for MMST II—there was a statistically significant 36% reduction in breast cancer mortality in the intervention groups (relative risk = 0.64; 95% CI: 0.45–0.89; *P* = 0.009). A harm-benefit analysis showed, however, that for every two breast cancer deaths prevented, one clinically insignificant cancer was diagnosed; for each breast cancer death prevented, 63 cancer-free women had been called back for further examinations; and for every 20 lives saved, one radiation-induced breast cancer death may have occurred. Recommendations for screening must therefore weigh mortality benefits against these negative effects. [Monogr Natl Cancer Inst 1997;22:63–67]

The conclusion in the publication of Malmö Mammographic Screening Trial (MMST) in 1988 was that invitation to mammographic screening may lead to reduced mortality from breast cancer, at least in women aged 55 years or over (*1*). When the code was broken after an average of 8.8 years of follow-up, there was no indication of an effect in women below age 55 at invitation. The accumulated breast cancer mortality was in fact higher in the invited group than in the control group.

The overview of the Swedish randomized trials (*2*), which was published in 1993, showed that invitation to screening was associated with a 24% statistically significant reduction in breast cancer mortality (95% confidence interval [CI]: 13%–34%). The 13% reduction in women younger than 50 years at invitation did not reach statistical significance (95% CI: −37% to 20%).

The design of the mammographic screening activities in Malmö following the end of MMST has allowed further estimates of the effect of screening in women below age 50. First of all, the controlled design in MMST, which included 7,984 women below 50 years of age born between 1927 and 1932, continued for four more years after the code was broken in 1988. In the present study, this cohort is called MMST I. Second, of the 17,786 women below age 50 who were born between 1933 and 1945, MMST cohort II, only 54% were randomly chosen to receive invitation to the screening that took place between 1978 and 1990. The remaining 46% were considered a control group. This group was invited to screening 1991 to 1994.

The effect of screening in women above age 50 takes several years to occur. We have for this reason chosen to base our estimate of the effect in women below age 50 on the accumulated breast cancer mortality in the two groups up until the completion of the first screening round for women in the control group.

Subjects

MMST I contained all women who were born between 1927 to 1932 and who lived in the city of Malmö from 1977 to 1978, and all women under age 50 at entry into original MMST. They were randomized to invitation on an individual basis, 50% being allocated to the control group (Table 1). The median age at entry was 47 years. The code was broken in 1988, when eight screening rounds had been completed. The controlled design for MMST I was continued up until the control group was invited in 1992. The first screening round for the control group was completed in 1993.

MMST II comprised all 17,786 women who were born 1933 to 1945 and were living in Malmö between 1978 and 1990. Of these, 53.9% were randomly allocated to receive invitation to screening. The plan was to invite these women when they turned 45, beginning in 1978. Due to limited resources, the plan could not be strictly adhered to, which means that, some years, no women could be invited, while other years two or even three birth-year cohorts were randomized and invited to examination. Seventy three percent of the women were 44 to 47 years of age

**Affiliations of authors:* I. Andersson, M.D., Department of Diagnostic Radiology, Malmö University Hospital, Malmö, Sweden; L. Janzon, Division of Epidemiology, Department of Community Medicine, Malmö University Hospital, Malmö, Sweden.

Correspondence to: Ingvar Andersson, M.D., Department of Diagnostic Radiology, Malmö University Hospital, S-205 02 Malmö, Sweden.

Table 1. Birth cohorts included in the mammographic screening evaluation in Malmö

			n	Accumulated no. of person-years at end of follow-up
MMST I Birth cohorts 1927–1932	First screening round 1977 to 1978	Invited group	3,954	61,069
		Control group	4,030	62,400
MMST II Birth cohorts 1933–1945	First screening round 1978 to 1990	Invited group	9,574	104,527
		Control group	8,212	81,636

when invited to the first screening round, the remaining being 47 to 48. The median age at entry was 46 years. The last birth-year cohort, women born in 1945, was invited in 1990. The first screening round of the control group took place between 1991 and 1994.

Methods

Mammography

State-of-the-art mammography was used throughout the trial. Two views, the craniocaudal and the oblique, were used as a baseline; subsequently, one (the oblique) or two were used, depending on the density of the parenchyma. With few exceptions, the interval between screenings was 18 to 24 months. Double reading was practiced when possible, but not consistently.

Women in MMST I were followed from date of first examination until death or December 31, 1993. Average follow-up time was 15.5 years. Women in the cohort MMST II were followed from date of first examination until death or December 31, 1995. Average follow-up was 10 years.

At the end of follow-up, nine screening rounds had been performed in MMST I and a mean of five rounds in MMST II. The attendance rate varied between 75% and 80%. In MMST I and MMST II, approximately 25% and 65% of the examinations, respectively, were performed before age 50.

Breast Cancer Mortality Surveillance

Breast cancer mortality was assessed by record linkage to the Swedish Cause of Death Register. The cause of death was validated by checking clinical records and autopsy reports when available. Breast cancer mortality is expressed as a percent and as deaths per 100,000 person-years of follow-up (Table 2). The effect of screening on breast cancer mortality is based on the pooled effect in MMST I and II. It is expressed in terms of relative risk (RR), with a 95% confidence interval around the point estimate.

Harm-Benefit Estimations

An attempt was made to illustrate the harm-benefit balance of mammographic screening. Seven variables were used to evaluate harm versus benefit. The three "positive" effects were prevented number of deaths, prevented number of cancers, and breast-conserving surgery; the four "negative" effects were false positive results, clinically insignificant cancer diagnosis, and risk of radiation-induced cancer (Table 3).

The number of prevented deaths was assessed by subtracting the number of deaths in the pooled invited group from the number in the pooled control group, and then adjusting for radiation-induced cancers. Furthermore, it was assumed that women with breast cancer who were prevented from dying of breast cancer did not develop metastatic disease.

One potential benefit associated with screening is breast-conserving surgery. To estimate the number of women who might undergo conservative surgery as a result of screening, we calculated the number of women with either stage I disease or stage II with tumor smaller than 3 cm and no engaged lymph nodes (T<3, NO) in the invited group and in the control group respectively. The difference was taken as a measure of the additional number of women that may have been offered breast-conserving surgery.

Since false positive results represent a potentially negative effect of screening, these were also assessed. A false positive result was defined as any classification of the findings at screening resulting in a recall for further work-up. The additional investigation was, in the majority of cases, additional mammographic images, sometimes supplemented by needle aspiration biopsy and, in a minority of cases, a surgical biopsy. The recall rate from screening for further examination was on average 4%.

A clinically insignificant cancer was defined as a cancer that would not have been diagnosed in the absence of screening. It is generally agreed that a proportion of ductal carcinoma *in situ* (DCIS) will not develop into invasive disease (*3*). It is also known that highly differentiated tubular carcinoma usually is a slow growing tumor with very good prognosis (*4*). The incidence of DCIS and highly differentiated tubular carcinoma in the invited and control groups was used to estimate the number of cancers that would not have been diagnosed in the absence of screening. Fifty percent of the difference was arbitrarily taken as an estimate of the percentage of clinically insignificant breast cancers detected by screening.

For the calculation of radiation-induced breast cancer death,

Table 2. Effects of screening on breast cancer mortality in women below 50

	Person-years of follow-up	Breast cancer mortality			RR	95% CI
		n	Percent	Per 10^5 person-years		
Invited groups (n = 13,528)	165,596	57	0.42	34.4	0.64	0.45–0.89; P = 0.009
Control groups (n = 12,242)	144,036	78	0.64	54.2	Reference group	

Table 3. Assessment of potential harm and benefit from screening women under 50 (per 100,000 person-years)

Positive effects	n	Negative effects	n
Prevented deaths	20	Further examination of false positives	1,260
Prevented cases of metastatic disease	20	Surgery for benign disease	56
Breast conserving surgery	36	Treatment of clinically insignificant cancer	10
Reassurance	?	Radiation induced breast cancer death	1
		False reassurance	?

the following assumptions were made: two views per breast, mean absorbed dose 2 mGy per view, biannual screening, 8% participation rate, and a linear dose-response relationship with an age-related risk (*5,6,7*).

Results

Effects of Screening on Breast Cancer Mortality

At the end of follow-up, the 13,528 invited women (in both MMST I and MMST II) had accumulated a total of 165,596 person-years. As Table 2 shows, 57 women in the invited groups died from breast cancer, corresponding to 34.4 per 100,000 person-years. The corresponding figures for the control groups were 78 breast cancer deaths or 54.2 per 100,000 person-years. This represents a statistically significant risk reduction of 36% (RR: 0.64; 95% CI: 0.45–0.89; $P = 0.009$). Figures 1 and 2 show the breast cancer mortality by year after entry. In MMST I, a lower mortality began to appear six years after entry in the invited group. In MMST II, the mortality curves had already started to separate two years after entry.

Harm-Benefit Evaluation

Table 3 shows that for every two breast cancer deaths prevented, one clinically insignificant cancer was diagnosed. For each breast cancer death that was prevented, 63 cancer-free women had been called back for further examinations. It was estimated that exposure to radiation may have induced one breast cancer death for every 20 that were saved.

Discussion

The follow-up of these two cohorts shows that repeated invitation to screening with state-of-the-art mammography was associated with a statistically significant 36% reduction of the

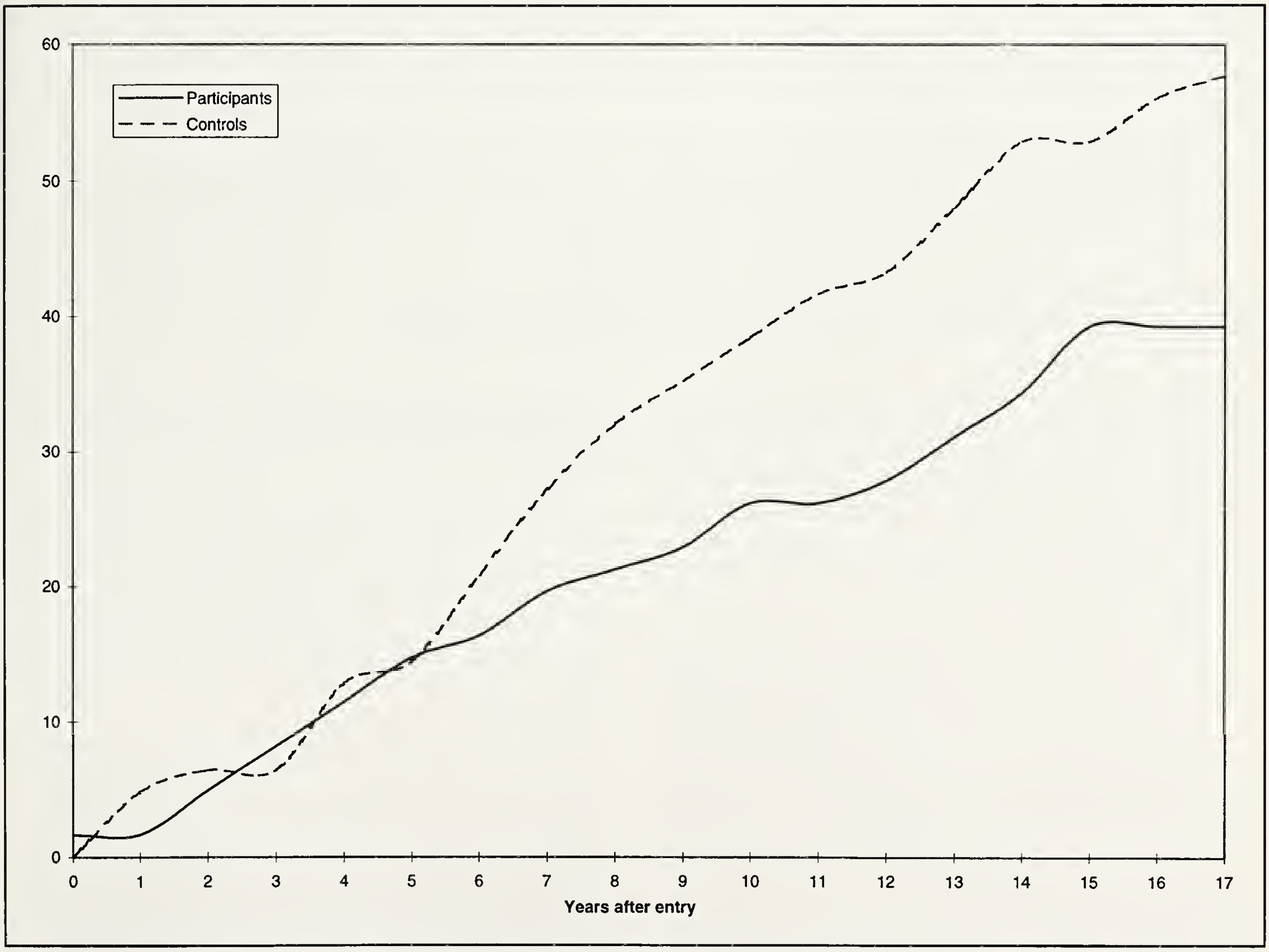

Fig. 1. Cumulated breast cancer mortality per 100,000 person-years in the invited and control group of MMST I.

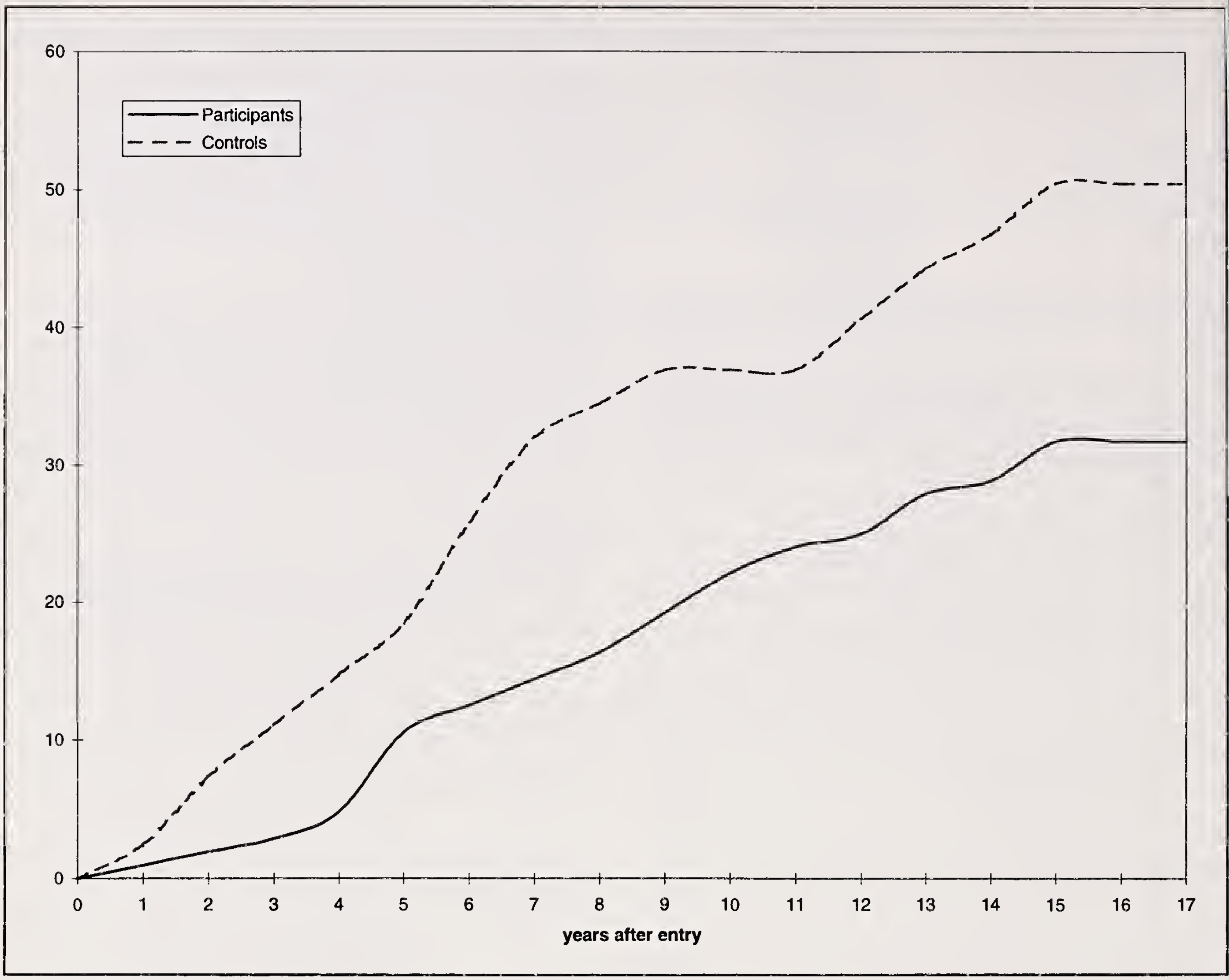

Fig. 2. Cumulated breast cancer mortality per 100,000 person-years in the invited and control group of MMST II.

mortality in breast cancer among women under 50 years of age at entry into the program. This result is in contrast to the outcome in the overview of the Swedish trials in which there was a nonstatistically significant reduction 8.8 years after entry (*2*). The longer follow-up in MMST I and MMST II, together with the greater number of screening rounds and better technology (compared to that of the late seventies and early eighties), certainly adds statistical power.

The results in MMST I and MMST II suggest that the effect on breast cancer mortality, in relative terms, is at least as good in women under age 50 as it is in those who are over that age. There is evidence to suggest that the progression of tumors in women under 50 is faster and the sojourn time therefore shorter (*8*). Accordingly, the screening interval should be shorter than in older women, possibly one year rather than the two-year interval that has been common. The current study shows, however, that even with an interval of 1.5 to 2 years, a significant effect on the breast cancer mortality can be achieved.

When considering a general recommendation of screening, one must consider the cost—both financially and in terms of life quality. Our harm-benefit analysis, however, did not consider financial cost, nor did we attempt to assess the potential positive and negative psychological effects. Furthermore, lives rather than life-years were used in quantifying positive as well as negative effects.

It is likely that the estimates of the number of false positive diagnoses and clinically insignificant cancers are conservative. In some programs, the rate of false positives has been twice as high (or more) as in the Swedish trials. This also applies to the detection rate of DCIS. The proportion of DCIS that will progress into clinically manifest disease is not clear. Estimates vary down to 30% or less (*3*). We chose 50% as an arbitrary estimate of that proportion.

It is also worth stressing that the significance of a mortality reduction expressed in relative terms is dependent on the underlying absolute risk. The current 36% reduction can also be expressed as a cumulative breast cancer mortality of 0.42% during the follow-up period in the invited group and 0.64% in the control group. This means that out of 10,000 invited women, 9,958 had not died form breast cancer in the invited group compared with 9,936 in the control group. Further, expressed in terms of the number needed to treat, these data imply that, on

average, 500 women had to undergo repeated screening for 12.5 years in order to save one woman from dying of breast cancer.

The radiation risk at low doses is hypothetical. However, most experts seem to agree that the risk cannot be completely ignored in women under age 50. Mammographic screening differs from most other radiologic examinations in that thousands of examinations of healthy women have to be performed to save one life. Furthermore, the total dose tends to increase the way screening mammography is practiced today, with two views instead of one, the use of a grid, a shorter screening interval, and probably the use of higher film density, even if modern film-screen technique has reduced the dose per image.

The continued follow-up of the mammographic screening activities in the city of Malmö lends support to the view that the relative risk reduction with regard to breast cancer mortality is similar in women below and above 50 years of age. The harm-benefit analysis indicates that the effect on mortality in premenopausal women may be associated with serious costs in terms of detection of clinically insignificant tumors, false positive findings, and even radiation-induced cancers.

References

(*1*) Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. BMJ 1988;297:943–8.

(*2*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.

(*3*) Frykberg ER, Bland KI. In situ breast carcinoma. Advances in Surgery 1993;26:29–72.

(*4*) Oberman HA, Fidler WJ. Tubular carcinoma of the breast. Am J Surg Pathol 1979;3:387–95.

(*5*) 1990 Recommendations of the International Commission on Radiological Protection. The International Commission on Radiological Protection. Pergamon Press, 1990.

(*6*) Sources and Effects of Ionizing Radiation. United Nations Scientific Committee on the Effects of Atomic Radiation. Unscear 1994 Report to the General Assembly with Scientific Annexes. United Nations New York, 1994.

(*7*) Land CE. Studies of cancer and radiation dose among atomic bomb survivors. The example of breast cancer. JAMA 1995;274:402–7.

(*8*) Tabar L, Larsson LG, Andersson I, Duffy SW, Nystrom L, Rutqvist LE. Breast-cancer screening with mammography in women aged 40–49 years. Int J Cancer 1996;68:693–9.

Variation in the Effectiveness of Breast Screening by Year of Follow-Up

*Brian Cox**

This meta-analysis assesses the effectiveness of breast screening by year from the start of screening for women aged 40–49 at study entry. Data from previous randomized controlled trials on breast cancer screening were combined, and cumulative and yearly breast cancer mortality rates and relative risks (RRs) were calculated for women offered screening compared to those not offered screening. At 7 years of follow-up, no reduction in breast cancer mortality from screening starting at ages 40–49 was found. At 10 years of follow-up, a nonsignificant reduction in breast cancer mortality was seen for women aged 40–49 at entry (RR = 0.93; 95% CI: 0.77–1.11). A nonsignificant excess of breast cancer mortality in those offered screening aged 40–49 was observed during the early years of follow-up in several trials. While the favorable effect of screening was observed within the first 5 years of study entry for women aged 50 or more, no similar effect was seen for women aged 40–49. The delayed effect for the 40–49 cohort may be attributable to 1) a biological difference in the effects of screening, which may be related to the onset of menopause, and 2) screening that occurred when women were aged 50 or more rather than before that age. [Monogr Natl Cancer Inst 1997;22:69–72]

The difference between younger and older women has been a surprise finding of the randomized controlled trials (RCTs) of breast cancer screening. For over a decade now various policy makers have noted this in their recommendations regarding breast screening. Recommendations have usually been made after careful review of the evidence and weighing the benefits and risks of screening, often by a range of independent experts in the field. Public health policy decisions, such as screening recommendations, require different ethical standards in assessing the balance of risks and benefits than the adoption of therapies or investigations in clinical practice (*1*). Most of the individual RCTs, however, were not specifically designed to examine the relative merits of starting screening at ages 40–49 compared to after age 50, and few studies entered sufficient numbers of participants aged 40–49 to examine this issue in detail. To provide an estimate of the effect of screening on breast cancer mortality, a meta-analysis of all RCTs was conducted, including an assessment of breast cancer mortality by each year after the start of screening for women offered screening compared to those not offered screening. Only the results from the RCTs were included, as these were less likely to be affected by the important biases of lead time, length bias, and selection bias in the assessment of screening effects. However, meta-analysis can conceal real differences that exist between studies and careful consideration of all trial results also should be undertaken.

Methods

Requests for aggregated numbers of breast cancer deaths and person-years for each year of follow-up by 5-year age group for women aged 40–49 at entry in both intervention and control groups were sent to the principal investigators of each of the RCTs except the HIP study, as this study is already completed and its results published (*2*). The investigators of the Edinburgh and Canadian trials kindly provided unpublished data to 10 and 14 years of follow-up respectively. Consequently, the analysis used earlier published data from some of the Swedish trials and more recent data from the Edinburgh and Canadian trials. A summary of the available data and their source is shown in Table 1.

The results of mammography screening of several types and frequency—with or without physical examination or breast self-examination—were combined for all 7 studies (*2–11*). For some studies, data were extracted from published tables or graphs (*12*). As published data from the trials were not routinely available in 5-year age groups, the analysis was performed for the women ages 40–49 at entry. The results of the two counties in the Swedish 2-county trial (S2C) were included separately. For some studies, only year of birth rather than the date of birth was available to determine age group. The design of the Canadian NBSS trial was slightly different from the other RCTs, since it was the only one specifically designed to examine breast screening in women aged 40–49; it assessed the *efficacy* of screening—that is, the mortality reduction for screened women—rather than the *effectiveness* of breast screening, the effect of providing a screening service for a defined population some of whom may not be screened. Despite considerable commentary and criticisms, which have been answered by the investigators, the study has not produced results considerably different from other studies at the same length of follow-up, and hence its results were included in this analysis.

A meta-analysis was undertaken to combine data for each individual year of follow-up and cumulatively by year of follow-up. The Gothenburg study was included in only the 7- and 10-year summary relative risks, since the trial results were not published and peer reviewed in full, and only summary data at 7 and 10 years of follow-up were available. These summary relative risks estimate the average effect of the intervention over the period covered adjusted for study size (*14*). Presence of major heterogeneity between studies was assessed to determine whether a random effects model was preferred in the meta-analysis;

**Correspondence to:* Dr. Brian Cox, Department of Preventive and Social Medicine, University of Otago Medical School, P.O. Box 913, Dunedin, New Zealand.

Table 1. Studies included: sources of data and contribution to the meta-analysis

Study	Age group	Reference	Years of follow-up	7-year summary	10-year summary	Yearly follow-up
Canadian NBSS	40–49	unpublished	14	yes	yes	yes
Edinburgh	45–49	unpublished	10	yes	ycs	yes
HIP	40–49	(2)	18	yes	ycs	yes
Malmö	45–54	(10)	10	yes	no	yes
	40–49	(4)		no	yes	no
Gothenburg	40–49	(4,13)	10	yes	yes	no
Swedish 2-county	40–49	(5)	12	yes	yes	yes
Stockholm	40–49	(4,11)	10	yes	yes	to 7 years

however, all analyses were able to be performed using a fixed effects model (*15*). Yearly mortality rates and cumulative mortality rates up to each year of follow-up were calculated separately for the intervention and control groups from the average yearly breast cancer mortality rates. This enabled estimation of the relative risk (RR) of breast cancer mortality among women offered screening compared to women not offered screening up to each year of follow-up. Approximate confidence intervals (CIs) were calculated using the normal approximation of the logarithms of the cumulative rates (*16*). Heterogeneity of the crude RRs associated with screening up to 7 years of follow-up between younger and older women was also evaluated (*17*). The results for younger women (those aged 40–49 [*2–6, 11*], 45–49 [*8,9*], or 45–54 [*10*]) and for older women (aged 50–59 [*3,4,7*], 50–64 [*2,8,9,11*], 50–69 [*5*] or 55–69 [*10*]) were calculated separately. In most analyses, results for the younger age group were calculated with and without Malmö subjects aged 45–54, since this trial included data from women aged 45–54 and not just those under age 50, but at 10 years of follow-up data for the women aged 45–49 alone was included in the calculation of the summary RR.

Results

At 7 years of follow-up, there was no significant reduction in breast cancer mortality in women under about 50 years of age (RR = 1.01; 95% CI: 0.80–1.28), with 144 deaths from breast cancer and nearly 670,000 person-years of follow-up among younger women offered screening and 132 deaths and over 604,000 person-years of follow-up among those not offered screening (Table 2). Exclusion of the Malmö trial reduced the summary RR to 0.98 (95% CI: 0.76–1.26). The results of the Canadian NBSS trial of women aged 40–49 were somewhat similar to the published results of other trials at an equivalent length of follow-up and were very similar to the results for Kopparberg county of the Swedish 2-county study over the first 7 years of follow-up (*12*).

In contrast, for older women at 7 years of follow-up, breast cancer mortality was reduced in those offered screening in all 7 trials, with 269 deaths from breast cancer among older women offered screening and 327 deaths among those not offered screening. The summary RR for older women at the 7-year follow-up was 0.74 (95% CI: 0.62–0.87), which differed little from the crude RR of 0.71 and was significantly different from the result in younger women.

At the 10-year follow-up (Table 2), with over 800,000 person-years of follow-up in the screened and nonscreened groups each, there was a nonsignificant reduction in mortality in women aged 40–49 offered screening, RR = 0.93 (95% CI: 0.77–1.11), with the Gothenburg trial (*13*) reporting a statistically significant reduction in breast cancer mortality among those offered screening. At both 7 and 10 years of follow-up, 3 studies showed higher, and 4 studies lower, average breast cancer mortality rates in those offered screening. In several trials, a nonsignificant excess of mortality from breast cancer was seen at short follow-up times. The overall cumulative breast cancer mortality rates for the intervention and control groups are shown in Figure 1. The cumulative breast cancer mortality rate for the intervention groups was similar to that for the control groups until more than 11 years of follow-up.

Table 3 shows the cumulative breast cancer mortality ratios estimating the RR of younger and older women offered screening compared to those not offered screening for each year of follow-up in 6 of the 7 studies (Gothenburg excluded). No consistent reduction in breast cancer mortality among younger women offered screening was seen, whether women aged 45–54 from the Malmö study were included or not. This was in marked contrast to the effect of screening by year of follow-up in older women where a benefit from screening appeared within the first few years of follow-up.

The breast cancer yearly mortality rate ratios estimating the RR during each year of follow-up, of those offered screening and those not offered screening for 6 of the 7 studies are also presented in Table 3. While overall breast cancer mortality appeared significantly higher in younger screened women during the third year of follow-up, this effect was reduced when women aged 45–54 in the Malmö study were excluded and was probably a chance finding, as multiple statistical comparisons were undertaken. For the yearly mortality rate ratios, data from the Stockholm study were only available for the first 7 years of follow-up, only 3 studies (the Swedish 2-county, HIP, and NBSS studies) contributed data through to 11 and 12 years of follow-up, and only two (HIP and NBSS) contributed through to the last 13–18 year follow-up interval. The test for linear trend in the rate ratio by year of follow-up was not statistically significant.

Table 2. Summary relative risks of breast cancer mortality at seven and ten years of follow-up

Duration of follow-up	Relative risk (95% CI)
7 years	1.01 (0.80–1.28)
10 years	0.93 (0.77–1.11)

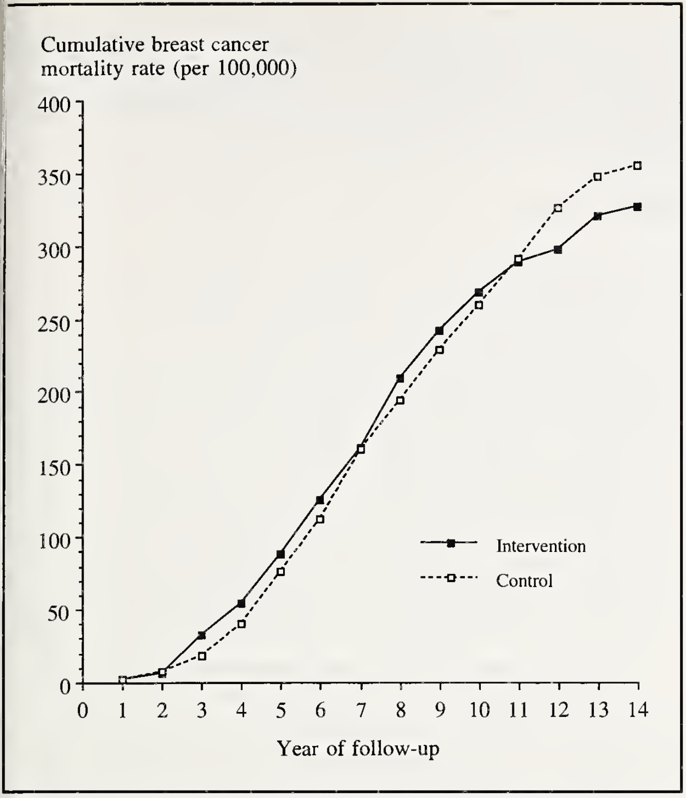

Fig. 1. Overall cumulative breast cancer mortality rates for intervention and control groups.

The overall pooled yearly breast cancer mortality rate ratio, adjusted by year of follow-up but unweighted for study size, was 0.97 (95% CI: 0.80–1.16). These results were not greatly altered when the Malmö study was excluded.

Conclusions

The combined results of all RCTs at about 10 years of follow-up with over 800,000 person-years of experience in each group did not show a significant reduction in breast cancer mortality among screened women.

The meta-analysis did not adjust for differences in screening frequency or for variation due to the cluster sampling used in the Swedish 2-county, Gothenburg, and Edinburgh studies. Adjustment for the cluster sampling method of subject selection has not been found to greatly alter the results in the Swedish 2-county study (*5*). Adjustment for socioeconomic status for the results of the Edinburgh study (*18*) or the use of a random effects model would not be expected to greatly alter the results of this meta-analysis but would slightly widen the CIs around the measure of effect. The design of the Canadian NBSS study was not considered to be sufficiently different from the other trials to warrant exclusion from the analysis, and it was the only trial that was specifically designed for women aged 40–49 and involved annual screening. Various reasons could be proposed for exclusion of many of the studies. For example, 61% of breast cancers detected among women aged 40–49 in the HIP study were detected by physical examination alone (*2*), which may have been the result of later presentation of clinical disease during the 1960s. Such exclusions were not considered to assist in the overall assessment of screening effectiveness in women aged 40–49 for the development of public policy.

It is important that the results of the different studies are combined at equivalent years of follow-up and not over many years of follow-up to avoid undue contribution from studies with the longest duration and to prevent attribution of a long-term effect to a more immediate one, especially where some variation in effect by year of follow-up is present. While the Gothenburg trial did not contribute to the cumulative mortality ratios for each year of follow-up, the results at 7 and 10 years of follow-up without the Gothenburg data (Table 3) were not markedly different from the 7- and 10-year cumulative RRs where all studies were included.

Reductions in breast cancer mortality among women initially offered screening at ages 40–49 have been reported at more than 10 years of follow-up (*13*), but such a delayed effect may well be due to screening they received after their 50th birthday, when

Table 3. Ratio of cumulative breast cancer mortality rates and yearly breast cancer mortality rate ratios for younger women and older women by length of follow-up for 6 studies (Stockholm, S2C, HIP, Malmö, Edinburgh, and Canada NBSS)

	YOUNGER WOMEN*			OLDER WOMEN*
Year of follow-up	Ratio of cumulative rates (95% CI)	Yearly mortality rate ratio† (95% CI)	Ratio of cumulative rates without Malmö	Ratio of cumulative rates
1	0.87 (0.12–6.16)	0.9 (0.1–6.2)	0.43	1.47
2	0.87 (0.28–2.69)	0.9 (0.3–3.5)	0.85	0.66
3	1.74 (0.92–3.31)	2.4 (1.1–5.4)	1.54	0.78
4	1.36 (0.86–2.15)	1.0 (0.5–2.0)	1.17	0.77
5	1.16 (0.82–1.63)	0.9 (0.6–1.6)	1.03	0.68
6	1.12 (0.84–1.49)	1.0 (0.6–1.8)	1.01	0.70
7	1.00 (0.79–1.28)	0.7 (0.5–1.2)	0.92	0.69
8	1.08 (0.87–1.35)	1.5 (0.9–2.5)	1.03	0.70
9	1.05 (0.86–1.30)	0.9 (0.5–1.6)	1.01	0.72
10	1.03 (0.85–1.26)	0.9 (0.7–1.7)	1.01	0.71
11	0.99 (0.82–1.20)	0.7 (0.3–1.4)	0.99	0.65
12	0.91 (0.75–1.11)	0.3 (0.1–1.2)	0.90	0.76
13–18	0.92 (0.75–1.13)	0.8 (0.3–2.0)	0.89	0.80

*Younger women aged 40–49 except when Malmö women aged 45–54 were included; older women were aged 50–69.
†Overall RR (unweighted) = 0.97 (95% CI: 0.80–1.16); without Malmö, RR (unweighted) = 0.94 (95% CI: 0.77–1.14).
Tests for linear trend (chi-square = 3.46, 1 degree of freedom, P = 0.06).
Chi-square for heterogeneity of relative risk with year of follow-up not significant.

it is known to be effective. Since the trials were not specifically designed to assess the relative value of screening starting at ages 40–49 compared to starting at age 50 or more, attribution of such a delayed effect to screening when women were aged 40–49 is difficult. Without clear evidence of benefit, the ethical foundation for starting breast screening at ages 40–49 is weak (*1*).

While analyses by age at diagnosis have been conducted, such analyses introduce lead time bias, as screen-detected cancers are by definition detected earlier—and therefore at younger ages—than other breast cancers. Without methods to adjust for this bias, an accurate estimate of any delayed reduction in breast cancer mortality from starting screening at ages 40–49 compared to starting at age 50 is not possible from current studies, and standard analyses will overestimate the size of any delayed benefit from screening starting at ages 40–49. This bias is only absent in the HIP study, which compared breast cancer mortality for women aged 40–45 at entry in the intervention and control groups who were diagnosed within 5 years of study entry. However, a detailed examination of the characteristics of breast cancer and age at diagnosis for women aged 40–49 at study entry who died of breast cancer at 9 or more years of follow-up may provide an estimate of the proportion of any reduction in breast cancer mortality due to initial screening at ages 40–49 rather than at age 50 or more. Nine or more years of follow-up is suggested, as there appears to be a consensus that no reduction in breast cancer mortality is consistently seen at earlier years of follow-up (*2*), and 9 years may be sufficiently long to reduce the effect of lead time bias in the comparison between those offered and those not offered screening.

The different effect in the two age groups over the years of follow-up suggests a biological difference in the effect of screening between women aged 40–49 and those aged 50 or more. The standard explanations for such an effect in screening, that of a different natural history or spectrum of disease for preclinical breast cancer in younger compared to older women, have been suggested and annual screening as a way of overcoming these has been advocated (*19*). This is an hypothesis that requires further research. The difference in starting screening at ages 40–49 rather than later may reflect a change in the effectiveness of breast screening about the time of the menopause, which would be consistent with other important differences in the epidemiology of breast cancer between pre- and post-menopausal women. Specifically designed studies, such as the UK trial and the proposed Eurotrial, are required to clarify this (*20,21*).

Acknowledgments

The assistance of Dr. Freda Alexander in providing data from the Edinburgh trial and Professor Anthony Miller in providing data from the Canadian trial is gratefully acknowledged. Helpful comments from Professors Mark Elwood and Anthony Miller and Dr. Ann Richardson on an unpublished manuscript upon which some of this article is based are also gratefully acknowledged.

References

(*1*) Cochrane AL, Holland WW. Validation of screening procedures. Br Med Bull 1971;27:3–7.

(*2*) Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer: the Health Insurance Plan Project and its sequelae, 1963–1986. Baltimore: Johns Hopkins University Press; 1988.

(*3*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993;341:973–8.

(*4*) Nystrom L, Larsson L. Breast cancer screening with mammography [letter]. Lancet 1993;341:1531–2.

(*5*) Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.

(*6*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. Can Med Assoc J 1992;147:1459–76.

(*7*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. Can Med Assoc J 1992;147:1477–88.

(*8*) Roberts MM, Alexander FE, Anderson TJ, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241–6.

(*9*) Alexander F, Anderson TJ, Brown HK, et al. The Edinburgh randomised trial of breast cancer screening: results after 10 years of follow-up. Br J Cancer 1994;70:542–8.

(*10*) Andersson I, Aspergren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. BMJ 1988;297:943–8.

(*11*) Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist L, Somell A. Randomized study of mammography screening-preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18:49–56.

(*12*) Elwood JM, Cox B, Richardson AK. The effectiveness of breast cancer screening by mammography in younger women. Online J Curr Clin Trials 1993; document 32, para 1-195.

(*13*) Smart CR, Hendrick RE, Rutledge JH, Smith R. Benefit of mammography screening in women ages 40 to 49 years. Cancer 1995;75:1619–26.

(*14*) Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.

(*15*) Pocock S. Editorial. Stat Meth Stat Res 1993;2:117–9.

(*16*) Clayton D, Hills M. Statistical Models in Epidemiology. Oxford: Oxford University Press; 1993:139.

(*17*) Breslow NE. Elementary methods of cohort analysis. Int J Epidemiol 1984; 13:112–5.

(*18*) Alexander F, Roberts MM, Lutz W, Hepburn W. Randomisation by cluster and the problem of social class bias. J Epidemiol Community Health 1989; 43:29–36.

(*19*) Tabar L, Fagerberg G, Chen H-H, et al. Efficacy of breast cancer screening by age. Cancer 1995,75:2507–17.

(*20*) Hakama M. Breast cancer screening with mammography [letter]. Lancet 1993;341:1531.

(*21*) Cox B. Benefit of mammography screening in women ages 40–49 years: current evidence from randomized controlled trials [letter]. Cancer 1996; 78:572–3.

The Quality and Interpretation of Mammographic Screening Trials for Women Ages 40–49

*Paul Glasziou, Les Irwig**

Using MEDLINE and the bibliographies of retrieved articles and reviews, we identified and systematically reviewed the quality and results of all randomized trials of mammographic screening that included women less than 50 years of age. Eight randomized trials were identified, 7 of which included women less than 50. Identified trials were assessed for the following design features: (a) method of randomization, (b) documented comparability of baseline data, (c) standardized criteria for breast cancer death, (d) blinded review of cause of death, (e) completeness of follow-up, and (f) use of an "intention to treat analysis." The quality of trials was generally high, with a total of almost 160,000 women randomized. In women aged 40–49 at entry, the overall, absolute risk difference between those invited and those not was 0.0004 (95% CI: 0 to 0.0009). Yet, what does this mean to a 40-year-old woman considering screening? If 10,000 women aged 40–49 years were screened regularly, then after a decade there would be about 4 less breast cancer deaths? Is that worthwhile? This is a difficult question, and it needs to be weighed against the problems arising from false positives and ductal carcinoma *in situ*. We recommend that women in this age group intending to be screened should be fully informed of these results in terms of absolute benefit. [Monogr Natl Cancer Inst 1997;22:73–77]

National committees in several countries have recommended that mammographic screening should commence at the age of 50. Since this is a somewhat arbitrary cut-off, many women and groups have asked, At what age should screening start? Unfortunately, the answer is not a simple yes or no, but involves a complex mixture of data interpretation, women's valuation of different outcomes, and resource implications. The major outcome of importance is death from breast cancer. If early detection did not result in a reduction in breast cancer deaths, then the only outcomes would be that women with screen-detected breast cancer would know about their cancer for a longer period, the false positive screens would have an unnecessary period of anxiety and investigation, and considerable resources would have been spent. Thus, we need to ask several questions when considering recommending across-the-board screening to women in their forties. First, we need to ask if there is adequate evidence of *additional* breast cancer mortality reduction from starting screening under age 50 compared with starting at age 50. Then, if there is an additional effect, we need to look at the absolute size of this benefit and compare it with the potential harms that come from screening: false positives screens, excessive treatment for those whose breast cancer would not have been detected before death, and the anxiety and costs. Finally, if women thought the benefit outweighed the harm, then the resource implications need consideration.

To determine if there is an additional benefit from starting earlier than age 50, women would ideally be randomized to either start screening earlier—at age 40, say—or to start screening at age 50. Unfortunately, none of the trials for which mortality data are available do this. Rather, they ask whether screening was better than no screening. This was appropriate given the lack of evidence for screening at the time the trials were commenced, but consequently these trials do not directly answer the question that is now being asked. Second, the ideal trial would have good long-term follow-up, as it would take many years to accumulate any benefit from earlier screening. Third, the trial would need to be extremely large in order to detect and reliably estimate the small benefits suggested to date. Finally, breast cancer mortality should be subject to a blinded and standardized evaluation of the cause of death. This ideal design, illustrated in Figure 1, is now being used in a trial that began in 1991 in the UK, with 150,000 of a proposed 195,000 women randomized thus far (S. Moss, personal communication); a similar multinational trial proposed by the UICC (International Union Against Cancer) (*1*) but involving 1,500,000 women and a pilot has also been started.

In the meantime, we need to try to answer the above questions as best we can, using the data available from previous screening trials. With this in mind, the aims of the present paper are 1) to examine the quality of the currently available trials of breast cancer screening compared to the ideal evidence, 2) to combine the currently available evidence to provide the best estimate of the absolute risk reduction in starting screening at 40 rather than 50, and 3) to ask whether the net benefit is worth the resource investment.

Methods

Quality of Trials

The major steps of a systematic review should involve (a) locating the appropriate studies, (b) critically appraising and

**Affiliations of authors:* P. Glasziou, Department of Social and Preventive Medicine, The University of Queensland, Brisbane, Australia; L. Irwig, Department of Public Health, University of Sydney, Camperdown, Australia.

Correspondence to: P. Glasziou, Department of Social and Preventive Medicine, The University of Queensland, Brisbane QLD 4072 Australia.

See "Note" following "References."

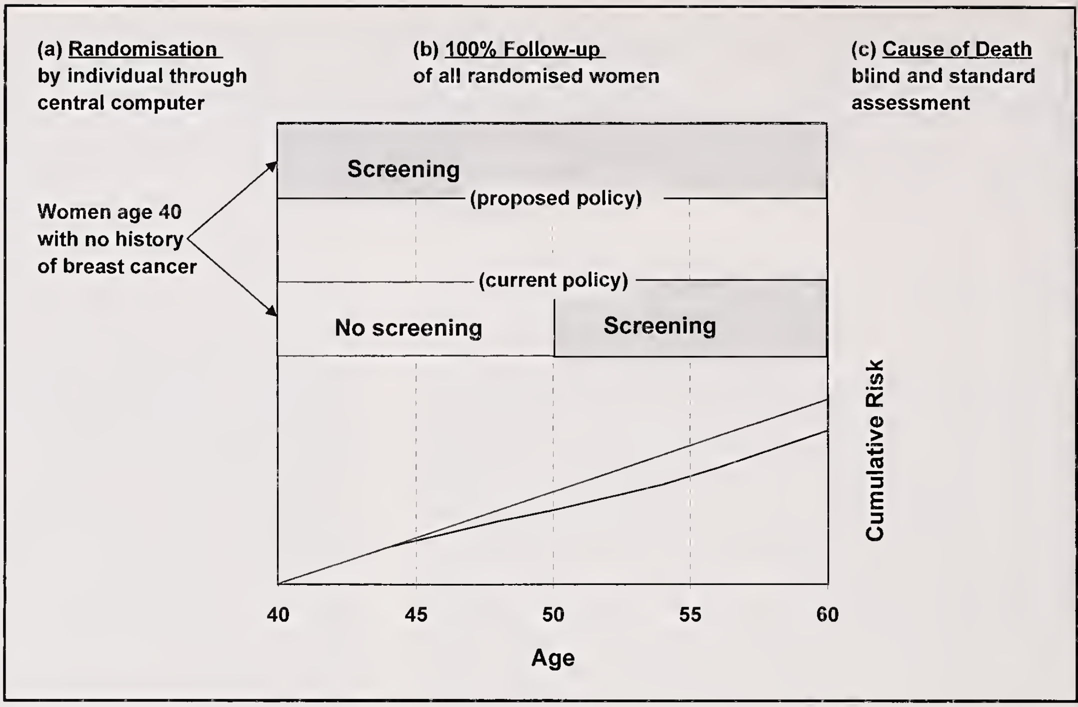

Fig. 1. The ideal trial to answer the question, "Should mammographic screening policy be extended to include the 40–49 age group?"

selecting the studies, and (c) analyzing and interpreting the results. Locating the studies of breast caneer screening is now straightforward, as they have been the focus of a number of systematic reviews (*2–6*). This seetion therefore focuses on critically appraising and selecting the studies; the next section will look at analyzing and interpreting the results of these studies.

Why is a systematic appraisal of study methods important? If trials are not assessed in an explicit and standardized process, there may be a tendency to apply different standards to different studies depending upon the results. For example, Mahoney (*7*) sent an invented paper to 28 reviewers asking them to appraise it in a number of categories. Fourteen copies were randomly allocated to have "positive" results; 14 were randomly allocated to have "negative" results. The papers were identical except for inversion of the results in figures and tables. The reviewers all found the paper topic to be highly relevant, but they were selectively critical of the *methods* in the paper with "negative" results. To minimize the bias created by this selective criticism of evidence, we eould use a standardized set of criteria for appraising the study's quality, or we could perform the critical appraisal "blind" to the study results by having a research assistant not involved in the appraisal process remove all references to results, or we could do both.

Table 1 shows a number of the quality features of the 7 available randomized trials that included women aged 40 to 49. An earlier blinded and standardized assessment of these trials (*6*) showed that all studies were of acceptably high quality. Notably, the much criticized Canadian National Breast Screening Study (*8*) ranked first, along with the Malmö trial, in methodological quality in this blinded review (Table 1). Since then, Schulz *et al.* (*9*) have published empirical data on which aspects of control trial design are most likely to cause significant bias; in particular, allocation concealment in randomization and the blinded assessment of outcomes were clearly demonstrated to be important.

In general, the trials listed in Table 1 had good randomization procedures. The Canadian study has been repeatedly criticized for a baseline imbalance in the numbers of advanced cancers. The number of cases of breast cancer detected at baseline physical examination with no nodes, 1–3 nodes, >3 nodes, or unknown were 35, 13, 17, and 0 respectively for the mammography group and 34, 16, 5, and 5 for the control group. Thus, the total cancers for each group are similar (65 versus 60), and those with and without nodes in each group are also similar (35 and 30 versus 34 and 21). This is unlikely to make an important prognostic difference, though a global test of the distribution of nodal status of cancers between the two groups is significant (chi-square [χ^2] = 11.9 on 4 degrees of freedom [df], P = 0.02) if we make no adjustment for the multiple comparisons possible on the Canadian baseline data—at least 9 comparisons were reported. There may have been a problem with cluster randomization in the Edinburgh study, as it appears to have noncomparable groups: there was a significant difference in non-breast cancer mortality (risk ratio = 0.80, P<0.0001) with a large portion of this difference explained by the baseline imbalance in socioeconomic status. Most of the other studies did not produce data on baseline equality in potential confounders. Since breast cancer incidence would be a major potential confounder for breast cancer mortality, we examined the numbers of breast cancers in all studies. Relative breast cancer instance is shown in Figure 2. We would generally have expected a slightly higher incidence in the screened group because of the lead-time; and, indeed, the overall increase is about 20%. This is seen in all studies (χ^2 for heterogeneity = 13.7 on 6 df; P = 0.03) except Gothenberg, where the incidence, surprisingly, is 8% lower in the screened group. However, this may be explained by the fact that the controls were screened at about 5 years after randomization. Nevertheless, the residual inequality between the two groups probably accounts for at least a portion of the better mortality reduction seen in the Gothenberg trial.

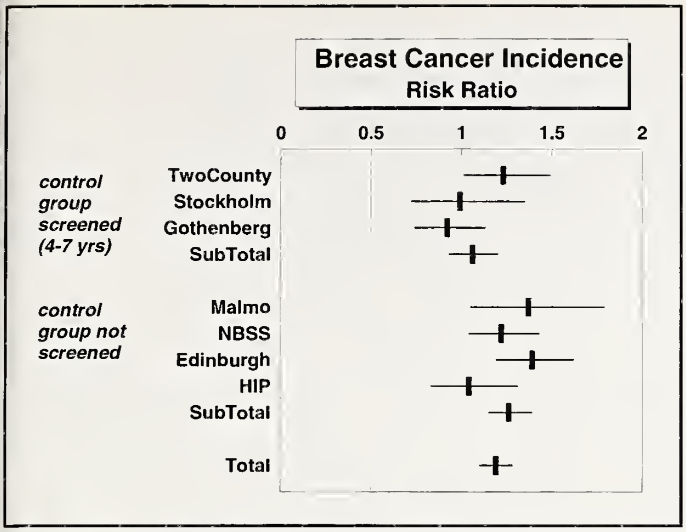

Fig. 2. Relative breast cancer incidence (proportion in screened/proportion in control). Balanced trials should generally show a small relative increase in incidence in the screened group.

All trials appear to have a high percentage of follow-up. Some trials, such as the Canadian study, ascertained vital status of all participants at the end of the study; most of the Swedish studies (*10–12*) did not explicitly do this, but given the excellent countrywide data tracking systems, all deaths (including breast cancer deaths) are likely to have been ascertained in these studies.

The three Swedish studies that did not use a blinded assessment of the cause of death have all had the cause of death re-reviewed in a blinded standardized fashion; however, this has made little difference to the overall results (*13*). Thus, 6 out of 7 studies (Edinburgh (*14*) is the exception) now have a blinded assessment of outcome.

Combining Studies

Which studies provide evidence on the incremental benefit of starting screening before rather than at the age of 50 years? None of the studies were explicitly designed to do this, but 3 of the studies did commence screening in the control groups a number of years after randomization. In the Two-County study, control women were screened between the fifth and seventh year; in the Gothenburg study, most women in the control group were screened between 4 and 7 years; in the Stockholm study (*12*), most women in the control group were screened at 4 years. Given that each of these studies included women aged 40 to 49, these figures mean that these 3 trials more closely approximate the ''ideal'' trial (randomized to screening starting at age 40 versus screening commencing at 50) than the other four studies. In the Edinburgh study, where the minimum age was 45, most women in the control group were screened between 5 and 11 years after randomization.

Which measure should be used to combine the results? The comparison of breast cancer deaths in the screened and control groups may be expressed in one of a number of ways. First, one can use ''relative risk,'' the relative ratio of the number of deaths in the screened groups compared with the control group. The problem with this method, however, is that both the relative risk (the cumulative risk ratio) and the hazard ratio (the instantaneous risk ratio) are likely to change over time, since (a) screening will take some years before benefits accrue and (b) if the control group has been screened, analysis that includes all breast cancer deaths will lead to some ''dilution'' because of the addition of deaths in those with screen-detected cancers beyond the age of 50 when both groups were being screened. To minimize the latter problem, the Swedish overview (*2*) used an ''evaluation'' model, which excluded breast cancer deaths from breast cancers detected after control group screening had commenced.

Second, one can use the final cumulative absolute risk difference in breast cancer deaths from starting screening earlier rather than at 50 (see Fig. 1)—that is, the difference between the proportion of deaths in the control group minus the proportion of deaths in the screened group. Under this model, including breast cancer deaths in the groups after screening has occurred in the control group is not a problem, but actually desirable—it more closely emulates the ideal of randomizing those screened at age 40 versus those at age 50. This was used in the ''follow-up'' model of the Swedish overview (*2*).

Thus, the better alternative is to use the final cumulative absolute risk difference. Furthermore, this is a more meaningful measure for assessing the clinical significance and in explaining the real benefits of screening to women and policy makers, as it includes the underlying risk of breast cancer death.

The results below have used the absolute risk differences weighted by the inverse of their variances. All calculations were done with MetaAnalyst software (J Lau, MetaAnalyst, version 0.988, Boston, 1996). The data include those presented at the recent conference in Falun (*15*) plus an interim update of the Canadian Trial (Cornelia Baines, personal communication).

Results

The relative risk from all 7 studies combined was 0.85 (95% CI: 0.71–1.01; 2P = 0.057)—that is, there was a 15% relative reduction in breast cancer mortality. This would seem to be about half the effect seen in the 50–64 age group. However, such a comparison does not account for the lower incidence and mortality risk in the younger age group. The method of expressing the results clearly makes a difference in their interpretation. In general, there is greater enthusiasm for results that are expressed as relative risks than those expressed as absolute risks (*16*). For example, Fahey *et al.* (*17*) have shown that health department officials are less enthusiastic about mammographic screening when results are expressed as absolute risk than as relative risk, with enthusiasm for number-needed-to-treat results falling between these.

The absolute risk difference results are shown in Figure 3. The trials here have been subgrouped by whether or not there was delayed screening in the control group (the former being desirable). For the 3 trials with delayed screening in the control group, the result was a risk difference of 4.0 per 10,000 with heterogeneity (χ^2 for heterogeneity = 1.6, P = 0.55). However, this subgrouping makes little difference to the overall results, and the absolute risk difference is in the range 4.0 to 4.2 per 10,000 in each of the two subsets and in all 7 trials combined. In both subsets and in the trials combined, the confidence intervals include a risk difference of zero—that is, no effect—and hence none of the results are statistically significant.

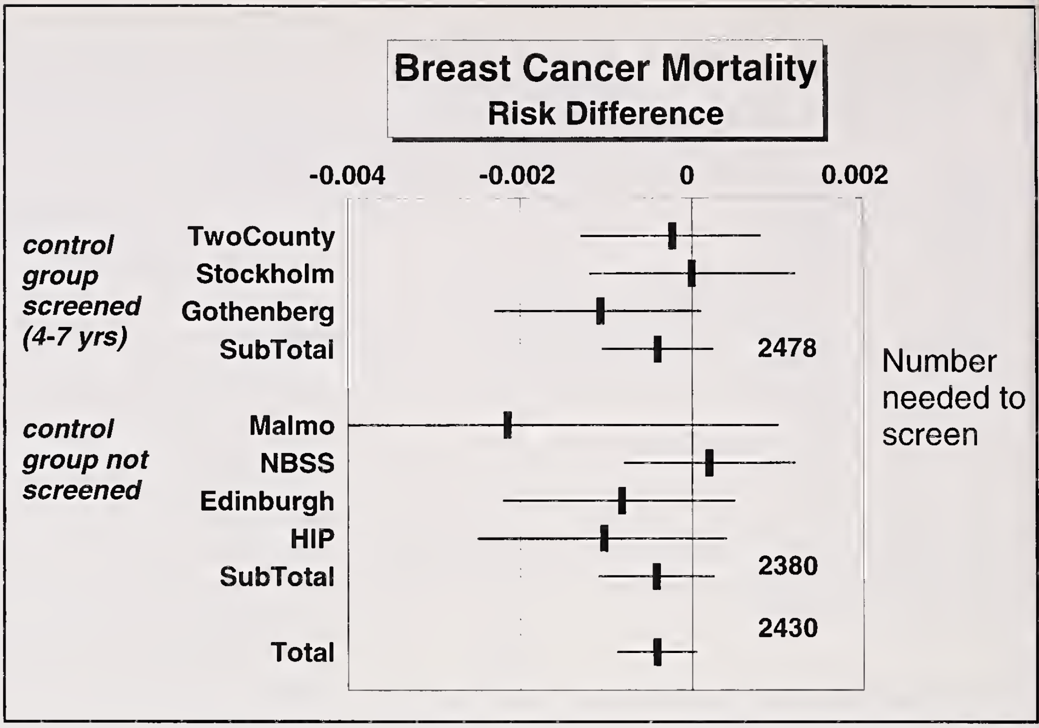

Fig. 3. Cumulative risk difference in breast cancer mortality (screened vs. control) for (a) studies which had delayed screening in the control group, (b) studies which had no screening in the control group, and (c) the two subtotals combined. Right-hand numbers are the numbers needed to screen.

An alternative method of presenting the results is as the "number needed to treat," which is the inverse of the risk difference. Thus, the above results can also be interpreted as meaning that 2,478 women would need repeated screenings in the age group 40 to 49 in order to prevent one breast cancer death about a decade later (95% CI: 951, infinite), based on the 3 studies with screening in the control group. The numbers are very similar for the 3 trials where the control group was not screened (2,380) and for all 7 trials combined (2,430; 95% CI: 1139, infinite).

Three other factors may influence the interpretation of these results. First, it must be remembered that there is some dilution from both nonattendance in the screened group and some screening in the control groups. A previous analysis (*18*) suggested that the attenuation reduced the "ideal compliance" effect by about one-third, but this attenuation was less in the 40–49 group, where noncompliance was less common. Second, the Edinburgh and Two-County studies were both cluster-randomized designs; however, this makes little difference, as the relative efficiencies are both around 90% (*18*). Finally, it should be noted that this analysis is effectively an age subgroup analysis of the combined screening trials, which include ages from 40 to 70 years. Overall, these studies clearly showed a statistically significant reduction in breast cancer mortality, and without showing heterogeneity by age, it is inappropriate to base conclusions purely on the statistical significance in a particular subgroup. A better alternative might be an empirical Bayes estimator (*19*), which would combine information from both the individual age subgroups and the all-age groups. However, this would require compatible (and preferably individual) follow-up data for all groups in all trials.

Conclusions

Seven randomized control trials for which mortality data are available have included women between 40 and 50 within their screening programs. However, no study was designed to test the current policy-relevant question of the incremental advantage of commencing screening at an age earlier than 50 versus commencing screening at age 50. Nevertheless, based on the substantial benefit seen for women over 50 and this analysis of the trials of women under the age of 50, there is good evidence of a small but real effect of mammographic screening in reducing the number of breast cancer deaths. About 2,500 women would need to be screened regularly in their forties in order to prevent one breast cancer death a decade later, though this will vary between populations and between countries, depending on the absolute risks of breast cancer mortality. For example, screening a population that has a 25% higher mortality from breast cancer than those in the trials will prevent roughly 5 deaths per 10,000 instead of 4 as in the present meta-analysis.

This benefit needs to be balanced against the harms and effort required by women undergoing screening, including the anxiety and investigation of false positives, and the perhaps unnecessary treatment of some women, such as those with ductal carcinoma *in situ* for whom the benefits are currently unclear. The next stage should be to inform the women involved in this decision about these results, to obtain their opinion as to whether this represents a net benefit, and to determine how strongly they would desire to participate. If women presented with this evidence expressed enthusiasm, then there would remain the societal question of whether the resources needed to be invested for this age group was seen as reasonable value for money.

References

(*1*) Multinational Breast Cancer Screening Conference Hosted by UICC in Geneva. UICC News 1993; No 4: December 1993.

(*2*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.

(3) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.
(4) Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.
(5) Elwood JM, Cox B, Richardson AK. The effectiveness of breast cancer screening by mammography in younger women [published errata appear in Online J Curr Clin Trials 1993; Doc No. 34 and 1994; Doc No. 121]. Online J Curr Clin Trials 1993; Doc No. 32.
(6) Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. Med J Aust 1995;162:625–9.
(7) Mahoney MJ. Publications prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research 1977;1:161–75.
(8) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years [published erratum appears in Can Med Assoc J 1993;148:718]. Can Med Assoc J 1992;147:1459–76.
(9) Schulz KF, Chalmers, I, Hayes RJ. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.
(10) Tabar L, Fagerberg G, Gad A, Baldetorp L, Holmberg L, Grontoft O, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985; 1:829–32.
(11) Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ 1988;297:943–8.
(12) Frisell J, Eklund G, Hellstrom L, Lindbrink E, Rutqvist LE, Somell A. Randomised study of mammographic screening — preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18:49–56.
(13) Nystrom L, Larsson L, Rutqvist LE, Lindgren A, Lindqvist M, Ryden S, et al. Determination of cause of death among breast cancer cases in the Swedish randomized mammography screening trials. A comparison between official statistics and validation by an endpoint committee. Acta Oncol 1995;34:145–52.
(14) Roberts MM, Alexander FE, Anderson I, Chetly U, Donnan PT, Forrest P, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241–6.
(15) Committee and Collaborators, Falun Meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun Sweden, March 1996. Int J Cancer 1996;68:693–9.
(16) Naylor D, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? Ann Intern Med 1992;117:916–21.
(17) Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. BMJ 1995;311:1056–59.
(18) Glasziou PP. Meta-analysis adjusting for compliance: the example of screening for breast cancer. J Clin Epidemiol 1992;45:1251–6.
(19) Davis CE, Leffingwell DP. Empirical Bayes estimates of subgroup effects in clinical trials. Control Clin Trials 1990;11:37–42.

Note

Cornelia Baines generously supplied information and answered questions about the interim update of the NBSS trial. We also thank the Australian National Breast Cancer Centre for providing funding support. Our thanks to the reviewers for helpful comments.

Efficacy of Screening Mammography Among Women Aged 40 to 49 Years and 50 to 69 Years: Comparison of Relative and Absolute Benefit

*Karla Kerlikowske**

In randomized controlled trials, screening mammography has been shown to reduce mortality from breast cancer about 25% to 30% among women aged 50 to 69 years after only five to six years from the initiation of screening. Among women aged 40 to 49 years, trials have reported no reduction in breast cancer mortality after seven to nine years from the initiation of screening; after 10 to 14 years there is a 16% reduction in breast cancer mortality. Given that the incidence of breast cancer for women aged 40 to 49 years is lower and the potential benefit from mammography screening smaller and delayed, the absolute number of deaths prevented by screening women aged 40 to 49 years is much less than in screening women aged 50 to 69 years. Because the absolute benefit of screening women aged 40 to 49 years is small and there is concern that the harms are substantial, the focus should be to help these women make informed decisions about screening mammography by educating them of their true risk of breast cancer and the potential benefits and risks of screening. [Monogr Natl Cancer Inst 1997;22:79–86]

Most experts agree that women aged 50 to 69 years should undergo screening mammography, since randomized controlled trials have shown screening mammography to reduce breast cancer mortality (*1,2*) and to be relatively cost-effective (*3,4*) for women in this age group. Whether or not recommendations should be extended to include screening starting at age 40 years remains controversial (*5–9*). This controversy stems from differences in interpretation of evidence and type of evidence used to evaluate whether screening mammography is efficacious.

Rationale for Using Evidence from Randomized Controlled Trials to Evaluate the Efficacy of Screening Mammography

In evaluating the controversy concerning routine screening mammography for women aged 40 to 49 years, it is important to remember that the goal of screening is to reduce the likelihood of death from breast cancer in a person who has the disease. Randomized controlled trials are the most unbiased means of assessing whether a screening test reduces the likelihood of death in a person who has the disease, and, for this reason, they are considered the gold standard when evaluating the efficacy of screening tests. In the randomized controlled trials of screening mammography, participants were randomly assigned to a screened or nonscreened (control) group to ensure that the screened and nonscreened groups were as alike as possible, so that any differences in outcome that were noted at the end of the trial could be ascribed to screening. In comparison, screening mammography programs and case series, which have no comparison group, are considered uncontrolled intervention studies and hence unsuitable for determining whether mammography decreases breast cancer mortality.

The debate concerning screening mammography among women aged 40 to 49 years has been perpetuated by reports from screening programs and case series claiming improved survival among younger women after initial breast cancer detection by mammography (*10–13*). Survival statistics favor screening since extra time is added to the interval between breast cancer detection and date of death by the fact that the diagnosis was made early. However, this lead-time in diagnosis may not affect date of death. For example, a 43-year-old woman may have breast cancer detected by mammography and a 45-year-old woman by finding a breast lump. If both women die of breast cancer at the age of 55, the former will have survived 12 years after the breast cancer detection and the latter 10 years. Although the 43-year-old woman lived an additional two years with breast cancer, having her breast cancer detected by screening mammography did not alter her life expectancy compared with the 45-year-old woman since both lived to be age 55. Thus, if survival statistics, rather than breast cancer mortality, are used as an endpoint to evaluate the benefits of mammography screening, it will appear as if screening is beneficial since the results will be unadjusted for time to diagnosis (i.e. lead-time bias).

Detection rates of early-stage cancer are also an inadequate measure of whether screening mammography decreases breast cancer mortality, since most cancers detected by mammography are primarily slow growing. If detection rates of early cancers are used as a surrogate endpoint for breast cancer mortality, it will appear as if screening is beneficial, since the results will be

**Affiliations of author:* Department of Epidemiology and Biostatistics, University of California, San Francisco, and General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco.

Correspondence to: Karla Kerlikowske, M.D., San Francisco Veterans Affairs Medical Center, General Internal Medicine Section, 111A1, 4150 Clement Street, San Francisco CA 94121.

See "Note" following "References."

unadjusted for rate of disease progression (length bias). Breast cancer is a heterogeneous disease, with some tumors growing relatively quickly, others so slowly that they may never cause breast symptoms, and yet others occurring somewhere in between. Within a year, fast-growing breast cancers may grow from undetectably small to large enough to cause symptoms, so that even annual screening may not detect the cancer—that is, it would be too small to detect by mammography on the first test and would already have become apparent before the next scheduled test. In addition, fast-growing tumors missed by screening are more likely to shorten a woman's' life substantially. Slow-growing breast cancers are more likely to be detected by screening mammography because they exist longer in an asymptomatic state. These slow-growing ones may have little or no impact on life expectancy. In addition, some small tumors detected by mammography metastasize early resulting in advanced stage disease at initial diagnosis (*14*). In this case, early detection may not be beneficial, even though the breast tumor was detected when it was relatively small.

If we knew the natural history of the various types of breast cancer, as well as their frequency and the length of time each existed in various growth states according to decade of age, it might be possible to correct for length and lead-time biases that are inherent in results from screening programs and case series. However, since this is not the case, only randomized controlled trials can provide an accurate picture of whether screening mammography and the treatment that follows decrease breast cancer mortality.

Results from Meta-Analyses of Randomized Controlled Trials

Meta-analysis is a quantitative approach for systematically combining results of previous research to arrive at conclusions about a body of research (*15,16*). Meta-analyses provide a more stable estimate of the effect of an intervention and put any one trial result into perspective by examining all similar trials.

There have been several meta-analyses published that combine data from randomized controlled trials in order to quantify the overall impact of screening mammography on breast cancer mortality (*1,17–19*). One of the earliest meta-analysis by Elwood *et al.,* used the fixed-effects Mantel-Haenszel statistical method to pool published data from six randomized controlled trials of screening mammography and found no reduction in breast cancer mortality in women aged 40 to 49 years seven years after the initiation of screening (*17*). A more recent meta-analysis combined data from eight randomized controlled screening mammography trials and found similar results (*1*). Four of the eight trials reported a nonsignificant increase in breast cancer mortality, whereas four reported a nonsignificant decrease, indicating a lack of statistically significant benefit or harm from screening mammography (Fig. 1). When data from the eight studies were combined using statistical methods described by Greenland (*20*) based on the assumption of fixed effects, the overall summary estimate showed a nonsignificant +2% (95% CI: −18% to +27%) increase in breast cancer mortality seven to nine years after the initiation of screening (Fig. 1).

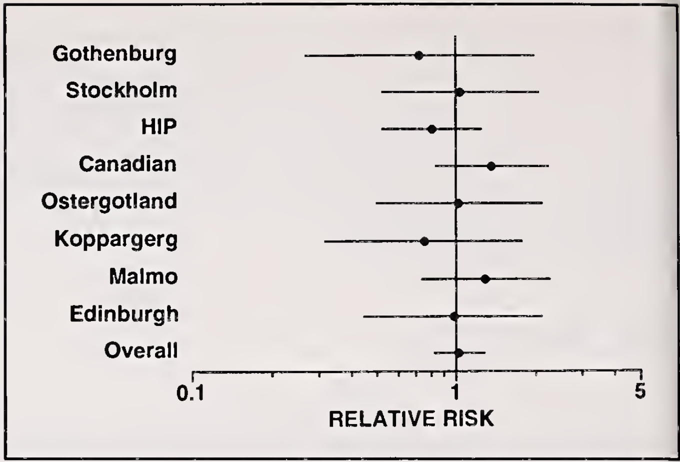

Fig. 1. Reduction in breast cancer mortality in women aged 40 to 49 years after seven to nine years of follow-up from the initiation of screening mammography among randomized controlled trials (adapted from reference (*1*)).

A separate meta-analysis, using a random-effects statistical method, combined results from the same eight randomized controlled trials and showed similar results with a nonsignificant breast cancer mortality reduction of −5% (95% CI: −23% to +18%) (*18*). Adjustment for cluster randomization in the Edinburgh trial and the Swedish Two-County trial did not affect the results (*18*). Importantly, despite the diverse study populations and interventions of the various screening mammography trials, the combined meta-analytic results of the eight randomized controlled trials were found to be homogeneous, indicating little variability of results between the individual trials (*1,17,18*). Taken together, the results from the three meta-analyses of the randomized controlled trials are consistent and indicate whether women aged 40 to 49 years underwent routine screening mammography or not, the risk of death from breast cancer was the same for the first seven to nine years after initiating screening.

One meta-analysis of data from randomized controlled trials has taken into account the various lengths of follow-up time after the initiation of screening (*1*). Combining trials with similar lengths of follow-up time is important, since trials with longer follow-up will have more breast cancer events and will be disproportionately weighted in meta-analyses, thus skewing results in favor of these trials. When published data for women aged 40 to 49 years reported from trials with at least 10 to 12 years of follow-up were examined, four of five studies had a relative risk estimate to the left of one, indicating a reduction in breast cancer mortality; however, all of the confidence intervals overlapped one (Fig. 2). When the five studies were combined using meta-analytic techniques (*20*), overall there was a trend toward a reduction in breast cancer mortality with an overall nonsignificant reduction of approximately −17% (95% CI; −35% to +6%) (*1*). Pooled data from the five Swedish trials (Fig. 3A), as well as results from the Health Insurance Plan (HIP) trial, also suggest an emerging benefit from screening mammography in younger women that does not occur for at least 9 to 10 years from the initiation of screening (*21–24*). If updated, unpublished results from the Gothenburg (*25*), Stockholm (*26*), Canadian (*27*), Malmö I and II (*28,29*), and Edinburgh trials (*28,30*) are combined with published results from the Kopparberg and Ös-

OMISSION

Monograph 22
Journal of the National Cancer Institute, 1997

Figure 3A, page 81 of the Monograph, "Efficacy of Screening Mammography Among Women Aged 40 to 49 Years and 50 to 69 Years: Comparison of Relative and Absolute Benefit," by Karla Kerlikowske, was not properly credited to its original publication, The Lancet. Monogr Natl Cancer Inst 1997;22:79–86

Figure 3A is an adapted version of the top graph in Figure 5, published in Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993;341:973–978.

Figure 3A is reproduced with permission from The Lancet Ltd.

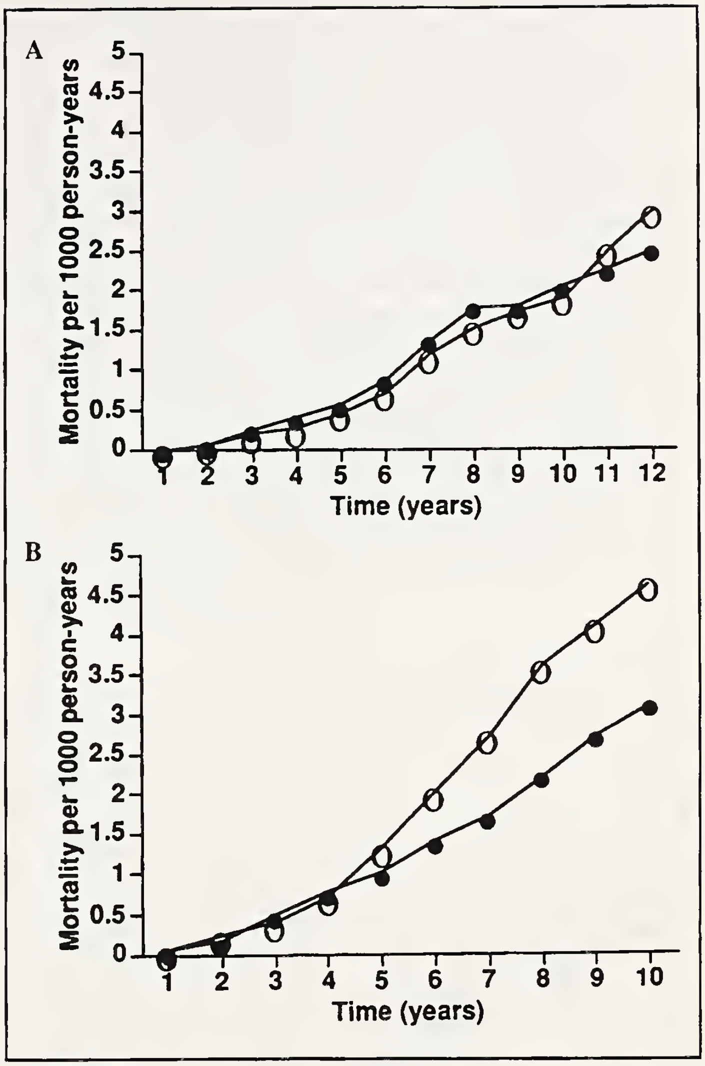

Fig. 3. Cumulative breast cancer mortality in screened and nonscreened women aged 40 to 49 years **(A)** [adapted from reference (*22*)] and women aged 50 to 69 years **(B)** [adapted from reference (*24*)]. ● = screened, ○ = nonscreened.

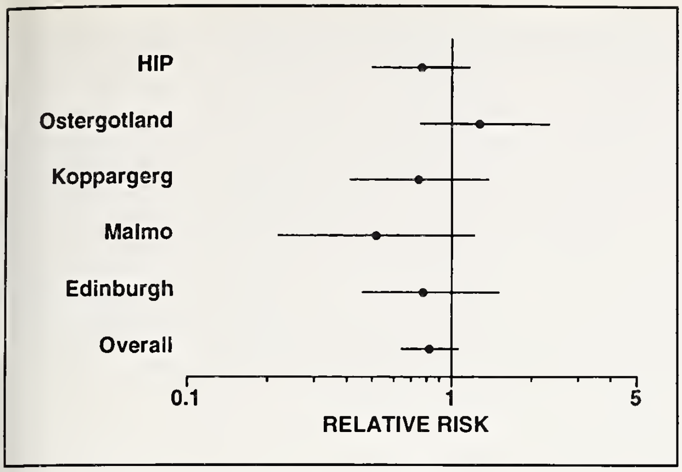

Fig. 2. Reduction in breast cancer mortality in women aged 40 to 49 years after 10 to 12 years of follow-up from the initiation of screening mammography among randomized controlled trials (adapted from reference (*1*)).

tergötland (*21*) and HIP trials (*23*) (Table 1) using the fixed-effects statistical method described by Greenland (*20*), the summary relative risk estimate shows a statistically significant −16% (95% CI; −29% to −1%) reduction in breast cancer mortality, similar in magnitude to an earlier report (*1*), 10 to 14 years after the initiation of screening (Fig. 4). Of note, a test for heterogeneity between study results was not significant (X^2 heterogeneity; $P = 0.4$), indicating that there was no statistically significant difference between the results of the individual studies.

One meta-analysis by C. R. Smart *et al.* (*19*), found contrasting results from other published overview analyses (*1,17,18*). Smart and colleagues reported a 24% reduction in breast cancer mortality among women aged 40 to 49 years who underwent screening mammography. Smart's' meta-analysis varied from other published meta-analyses because results were combined from studies with a wider range of follow-up times (7 to 18 years), unpublished data from the Gothenburg trial were included (*28*), and results from the Canadian National Breast Screening Study (*31*) were excluded. As demonstrated above, not stratifying results by length of time from initiation of screening disguises the fact that if screening mammography is effective in women aged 40 to 49 years, its effectiveness only appears 10 years after the initiation of screening (Fig. 3A). Meta-analysts are encouraged to consider unpublished data to avoid publication bias, but the drawback to that is, since the findings have not been peer reviewed, they may contain errors and inconsistencies. For example, it is puzzling that the Gothenburg trial, whose study methods have never been published, is the only randomized controlled trial that shows a greater benefit for screening women in their forties than for screening women aged 50 and older (*1*). Smart omitted the Canadian National Breast Screening Study (*31*) from his meta-analysis, claiming that, since the study population consisted of volunteers rather than being population-based, it should not be combined with the other trials (*19*). This seems to be a relatively weak criterion for study exclusion, since it is not obvious that having volunteers as study participants would make it more or less difficult to find a reduction in breast cancer mortality among screened women.

In order to minimize selection bias in performing a meta-analysis, it is important that all similar trials are combined. Each of the randomized controlled trials listed in Table 1 is slightly different and could be excluded from a meta-analysis of randomized controlled trials of screening mammography for some aspect of its study design or intervention: some trials, for instance, used one-view mammography instead of two-view mammography, which is considered optimal for women aged 40 to 49 years; others used biennial rather than annual screening, also considered optimal for women aged 40 to 49 years; and others combined clinical breast exam with mammography, making it difficult to assess the independent contribution of mammography. Despite these differences, the confidence intervals for all of these studies overlap each other (Fig. 4), indicating the results from these studies are not greatly dissimilar and can be combined to summarize the results. Thus, it is not methodologically appropriate to selectively omit any one trial, and doing so may introduce selection bias into the results. If adjustment for length of follow-up, data inconsistencies (*32*) and selective study exclusions are taken into account, Smart's' results are similar to those previously published (*1*).

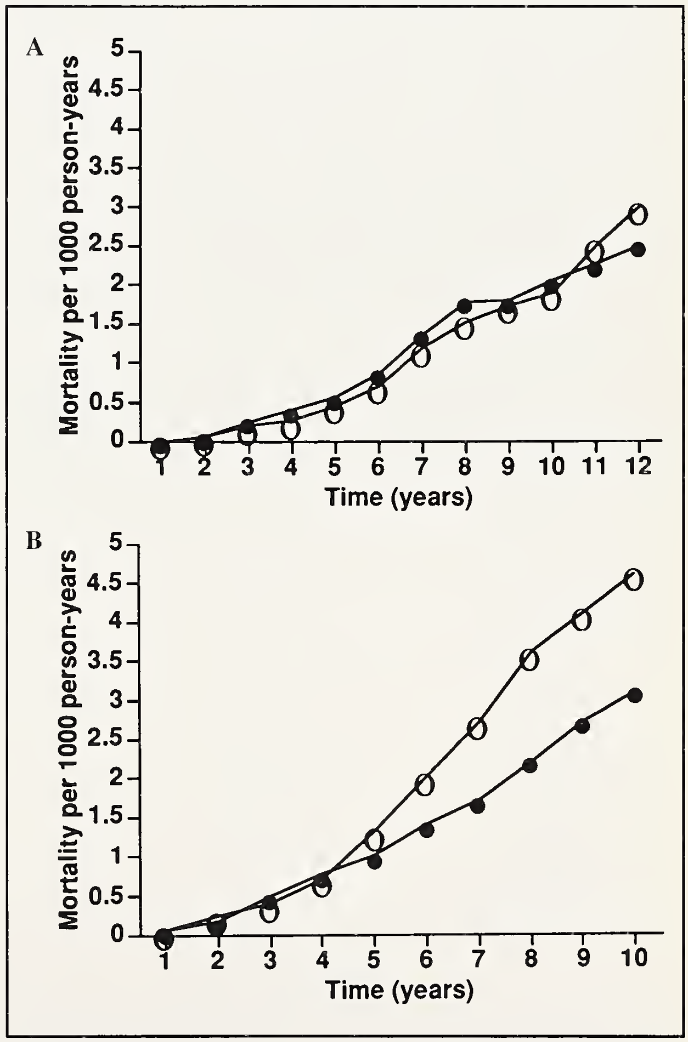

Fig. 3. Cumulative breast cancer mortality in screened and nonscreened women aged 40 to 49 years **(A)** [adapted from reference (*22*)] and women aged 50 to 69 years **(B)** [adapted from reference (*24*)]. ● = screened, ○ = nonscreened.

Table 1. Randomized controlled trials included in updated meta-analysis for women aged 40 to 49 years

Study (ref)	Start date	Ages (yr.)*	Screening interval (mo.)	# of mammographic views	Annual clinical breast exam	Duration of follow-up (yr.)	Relative risk (95% CI)
Gothenburg (*25*)	1983	39–49	18	2	no	12	0.56† (0.32–0.98)
Stockholm (*26*)	1981	40–49	24–28‡	1	no	11.4	1.08† (0.54–2.17)
HIP (*23*)	1963	40–49	12	2	yes	10	0.77 (0.50–1.16)
Canadian (*27*)	1980	40–49	12	2	yes	10.5	1.14† (0.83–1.56)
Östergötland (*21*)	1977	40–49	24	1	no	13	1.02 (0.52–1.99)
Kopparberg (*21*)	1977	40–49	24	1	no	13	0.73 (0.37–1.41)
Malmö I (*28*)	1976	45–49	21	2	no	14	0.67† (0.35–1.27)
Malmö II (*29*)	1978	45–48	21	2	no	12	0.69† (0.44–1.09)
Edinburgh (*28,30*)	1978–82§	45–49	24	2¶	yes	10–14	0.73† (0.43–1.25)

*Age range of participants at start of mammography screening.
†Data presented but unpublished in peer-reviewed journal.
‡First round 28 months after baseline exam, second round 24 months after first round.
§Initial randomization 1978; additional women aged 45–49 years randomized starting in 1982.
¶First round, two-view mammography; subsequent rounds, one-view mammography.

Are Data from Randomized Controlled Trials Conclusive?

Some have argued that it is inappropriate to use meta-analytic techniques to pool data to evaluate the efficacy of screening mammography among women aged 40 to 49 years; that such subgroup analyses are inappropriate when initial screening trials were designed for women aged 40 to 74 years (*9*). However, this is exactly the purpose of a meta-analysis: to combine data from several trials to obtain a more stable estimate of the effect of an intervention when there are insufficient numbers of subjects in any one trial to yield a meaningful conclusion (*15,16*). If subgroup analyses by age at initiation of screening are to be discounted, then consideration must be given only to the sole randomized trial specifically designed to address the efficacy of screening mammography in women aged 40 to 49 years, and this trial has yet to show a reduction in breast cancer mortality among screened women (*27,31*).

Others have argued that the randomized controlled trials of screening are methodologically flawed and should not be used to conclude that mammography is not beneficial for women aged 40 to 49 years. Yet, results from these same trials are used to support mammography screening among women aged 50 to 69 years. A meta-analysis (*1*) of data in women aged 50 and older from eight randomized controlled screening mammography studies demonstrated an overall significant 27% (95% CI: −37% to −6%) reduction in breast cancer mortality after seven to nine years from the initiation of screening (Fig. 5). Of note, despite differences in types of randomization (cluster, individual), interventions (screening intervals from 12 to 33 months, single-view or two-view mammography, screening with or without clinical breast examination), and study populations, screening mammography trials have consistently demonstrated a reduction in breast cancer mortality among screened women aged 50 to 69 years.

Screening mammography trials are also criticized for using obsolete technology, implying that modern mammography has an increased ability to detect breast cancer in younger women. Several published studies, however, show that the sensitivity of modern mammography, in particular its sensitivity to deteet invasive cancer, is still lower for women less than age 50 than for women aged 50 and older, despite improvements in technology (*33–38*). Still others have argued that screening would be effective in younger women if the interval between each mammographic examination were one year rather than two years (*39*). Only two trials have screened women aged 40 to 49 years an-

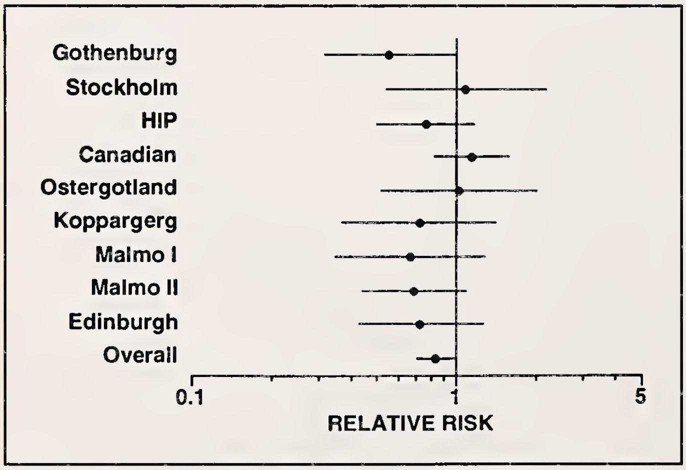

Fig. 4. Updated results of reduction in breast cancer mortality in women aged 40 to 49 years after 10 to 14 years of follow-up from the initiation of screening mammography using published and unpublished results from randomized controlled trials.

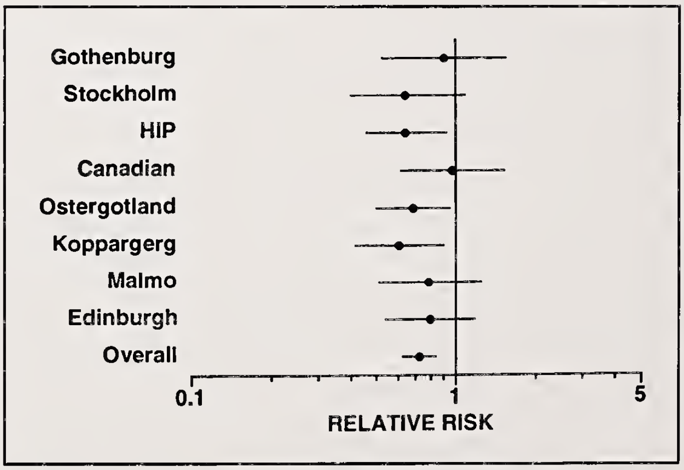

Fig. 5. Reduction in breast cancer mortality in women aged 50 to 74 years after seven to nine years of follow-up from the initiation of sereening mammography among randomized controlled trials (adapted from reference (*1*)).

nually (*23,31*), and there was variability in their findings. The HIP trial showed a nonsignificant reduction in breast cancer mortality among women in the group eligible for screening nine years after the initiation of screening, whereas the Canadian trial found a nonsignificant increase seven years after the initiation of screening. Among women aged 50 and older, whether they are screened annually or biennially, the reduction in breast cancer mortality is the same—that is, more frequent screening does not result in more deaths prevented (*1*). Therefore, given the differences in tumor biology among younger women, it is optimistic to think that more frequent screening in younger women will necessarily result in the same benefit that is evident in older women. Screening more frequently than every two years will, however, increase the number of unnecessary diagnostic evaluations, the detection of cancers of low malignant potential, and the cost of screening (*33*).

Lastly, proponents of screening mammography contend that randomized controlled trials have enrolled too few women to demonstrate a statistically significant benefit from screening mammography among younger women. If the explanation was merely lack of statistical power, and the efficacy of screening mammography in younger women was similar to that in older women, then a reduction in breast cancer mortality should begin to appear after four to five years from the initiation of screening, as in women aged 50 to 69 years (Fig. 3B), and should become statistically significant with longer follow-up, that is, the percentage reduction in breast cancer mortality observed at seven to nine years from the initiation of screening among women aged 40 to 49 years should be similar to that reported at 10 to 12 years, but with wider confidence intervals around the point estimate. This does not appear to be the case, since the data do not show a gradual separation of the mortality curves between screened and nonscreened groups (Fig. 3A). In fact, the data show slightly higher breast cancer mortality among screened women the first 10 years after the initiation of screening. Arguing that too few women have been enrolled to demonstrate a statistically significant benefit from screening mammography underscores that breast cancer is not as common in younger women as in older women and that mammography is not as effective in reducing breast cancer mortality in younger women.

In summary, the evidence from pooled results of randomized controlled trials may be interpreted in one of two ways: First, results from meta-analyses provide evidence, even if with low power, that screening younger women provides no benefit the first seven to nine years from the initiation of screening; however, a trend toward reduced mortality emerges after 10 years that appears to be smaller than that observed in older women; or second, results from meta-analyses are collectively inadequate, since these analyses are based on retrospective subgroup analysis. In either case, the scientific evidence to support mass mammography screening for women aged 40 to 49 years is not compelling.

Why Is the Benefit Among Younger Women Delayed?

Although pooled results of large randomized controlled trials failed to demonstrate any benefit in women aged 40 to 49 years after seven to nine years of screening (*1,17–18*), some have argued that the trend toward a reduction in breast cancer mortality that begins after 10 years of screening should not be ignored (*5*). It is unclear why any potential benefit from screening mammography in women aged 40 to 49 years should be delayed a decade. It could be that some of the breast cancers detected among women who start screening at ages 40 to 49 years are actually detected at or after age 50, when mammography is known to be efficacious. The HIP trial has published screening results by age at detection, and it found that 85% of breast cancers in women who started screening between ages 40 and 49 were diagnosed between ages 45 and 54. Almost all of the decrease in breast cancer mortality among women eligible for screening aged 45 to 49 years at entry in the HIP trial occurred in those who had breast cancer detected at ages 50 to 54 years (*40*). Furthermore, the majority of women in the Edinburgh and Malmö trials, which also showed no benefit seven to nine years from the initiation of screening but a trend toward a delayed benefit after 10 to 12 years (*1,2*), were also probably aged 50 or older when their breast cancer was diagnosed, since the youngest age of women at the start of screening was 45 years old. The same rationale has been applied to the Swedish data, since women who started screening at ages 40 to 49 years were offered regular screening mammography with many actually being 50 or older in the ensuing years. Computer modeling of the Swedish breast cancer screening trial data has also suggested that some of the observed decrease (about 30–40%) in breast cancer mortality for women aged 40 to 49 years at trial entry may be attributable to continued screening after women reach age 50 (*41,42*).

Why is mammography efficacious as early as four to five years after the initiation of screening in older women? One explanation is that, among women aged 50 and older, the sensitivity of mammography to detect invasive cancer is relatively high, resulting in few undetected cancers. This relatively high sensitivity is probably due to two factors: a greater proportion of older women tend to have fatty breast density, which allows easy detection of breast cancer; and tumor growth rates are not as rapid as in younger women, allowing sufficient time for detection of small tumors (*33,43*). Thus, among women aged 50 and older, mammography detects the majority of tumors and detects them when they are more curable than if they were detected clinically. In contrast, the sensitivity of screening mammography to detect invasive breast cancer is lower among women aged 40 to 49 years compared to women aged 50 and older (75% versus 93%) (*33*). Conventional thinking has been that this lower sensitivity is due to younger women's' breasts being more radiographically dense. However, only two studies have evaluated the sensitivity of mammography according to radiographic breast density, and both found that breast density did not influence the sensitivity of mammography in women less than 50 years of age (*21,33*). An alternative explanation is that a greater proportion of invasive breast cancers are aggressive in younger women and therefore grow more rapidly, resulting in more interval cancers between regular screening examinations. This theory is supported by the observation that the sensitivity of screening mammography decreases with increasing tumor size. That is, tumors that are not detected by mammography are larger at clinical presentation than tumors that are mammographically detected. A lower sensitivity for detecting large tumors is more

marked in younger than in older women, suggesting that tumors not detected by mammography in these younger women are especially rapid growing (*33*). This is further supported by the finding that the sensitivity of mammography decreases rapidly as the length of time between screenings increases (*33,44*), and by the observation that, among women aged 40 to 49 years, a greater proportion of small tumors detected by screening mammography are associated with positive lymph nodes as compared with older women (*14,45*). Consequently, among women aged 40 to 49 years, the proportion of slow-growing tumors with a good clinical prognosis detected by screening mammography is probably small, which may account for both the marginal and delayed benefit from screening observed in randomized controlled screening mammography trials. Taken together, these findings suggest that the tumor biology is different in younger than in older women and that the small, delayed benefit observed in the randomized controlled trials for women aged 40 to 49 years may be more of a reflection of the biology of the tumor than of screening mammography.

If the delayed reduction in breast eancer mortality is primarily due to detection of indolent tumors among younger women, such as slow-growing invasive tumors or ductal carcinoma *in situ*, some of these slow-growing tumors could well be detected satisfactorily at or after age 50 years, providing the same reduction in risk of breast cancer deaths as if the tumors were detected in their forties. If the delayed reduction in breast cancer mortality is, in part, because some of the breast cancers detected among women who start screening at ages 40 to 49 years are actually detected at or after age 50, this is further evidence that starting sereening at age 50 is reasonable.

Absolute Benefit

Reporting the relative risk reduction in breast cancer mortality among women undergoing screening mammography compared to those who do not is not as clinically relevant as reporting the absolute risk reduction due to screening. Reporting the relative risk reduction between screened and nonscreened populations as a percentage obscures differences in the incidence of disease among populations. This is particularly important when the incidence of disease events (e.g., breast cancer deaths) is low, as is the case for women aged 40 to 49 years. The absolute risk reduction or risk difference (difference in risk of dying of breast cancer between screened and nonscreened women) takes into account the underlying incidence of disease events and expresses how much the risk of death from breast cancer is reduced by screening. The reciprocal of the absolute risk reduction is the number needed to screen to prevent one death (*46*). The number needed to screen is a measure of clinical significance that allows comparison between groups with differing underlying incidence of disease events and quantifies the effort required by patient and physician to prevent one death.

A Markov simulation model that takes into account competing causes of death has been used to determine the number needed to screen to prevent one death if women are screened biennially from ages 50 to 69 years, and the number needed to screen to prevent one death if screening was extended to included annual screening every one to two years for women ages 40 to 49 years (*47*). Assuming that mammography screening among women who initiated screening at age 50 results in a 27% (*1*) reduction in breast cancer mortality starting five years from the initiation of sereening, it has been estimated that 270 fifty-year-old women would need to be screened biennially for 20 years to prevent one death. This means approximately 2,700 screening mammographic examinations would need to be performed to prevent one death (*47*). Assuming that all of the delayed benefit in breast cancer mortality among women who initiated screening at age 40 results from detecting cancer before age 50 and that the delayed reduction is at least 16% starting 10 years from the initiation of screening, it has been estimated that 2,500 forty-year-old women would have to be screened every one to two years for 10 years to prevent one death (*47*). This means between 12,500 and 25,000 screening mammographic examinations would have to be performed to prevent one death. The tenfold difference between younger and older women in the number needed to screen to prevent one death is due to the lower incidence of breast cancer among women aged 40 to 49 years, the delay in benefit from screening and the lower relative risk reduction in breast cancer mortality from screening mammography. If the delayed reduction in breast cancer mortality was as large as 27%, it would still require performing between 7,150 and 14,300 screening examinations on women aged 40 to 49 years to prevent one death (*47*). Therefore, even assuming an optimistic reduction in breast cancer mortality from screening mammography, the number needed to screen and the total number of mammographic examinations needed to prevent one death is very large for women aged 40 to 49 years.

Conclusion

In summary, based on the results of meta-analyses, there is no reduction in breast cancer mortality seven to nine years after the initiation of screening among women aged 40 to 49 years who undergo screening mammography. There appears to be a delayed reduction in breast cancer mortality 10 years after the initiation of screening, and a proportion of this reduction is benefiting women aged 50 to 59 years rather than women in their forties. It is important to emphasize that if screening mammography is effective in reducing breast cancer deaths among women aged 40 to 49 years, the reduction in deaths does not occur for at least a decade following the initiation of screening and appears to be smaller than the reduction observed in women aged 50 and older. Given that the incidence of breast cancer for women aged 40 to 49 years is lower and the potential benefit from mammography screening smaller and delayed, the absolute number of deaths prevented by screening women in this age group is likely to be much less than by screening women aged 50 and older.

Many people feel that it is acceptable to perform widespread screening mammography in women aged 40 to 49 years despite lack of compelling evidence of benefit, yet proven associated risks (*5,39,48–50*). In the case of screening mammography, these risks include additional diagnostic evaluations and the associated morbidity and anxiety, the potential for detecting and surgically treating clinically insignificant breast lesions, and the potential false reassurance resulting from having a normal examination (*51*). Before making a blanket recommendation to all healthy women in an age group to have a screening test, the

benefits of the intervention should be proven and should clearly outweigh the risks (*52–54*). Because the absolute benefit of screening mammography for women aged 40 to 49 years is small and there is concern that the harms are substantial (*55–58*), the focus should be to help these women make informed decisions about screening mammography by educating them of their true risk of breast cancer and the potential benefits and risks of screening (*59*).

References

(*1*) Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.
(*2*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.
(*3*) Eddy DM. Screening for breast cancer. Ann Intern Med 1989;111:389–99.
(*4*) Kattlove H, Liberati A, Keeler E, Brook RH. Benefits and costs of screening and treatment for early breast cancer. Development of a basic benefit package. JAMA 1995;273:142–8.
(*5*) Sickles EA, Kopans DB. Mammographic screening for women aged 40 to 49 years: the primary care practitioner's' dilemma. Ann Intern Med 1995; 122:534–8.
(*6*) Harris R, Leininger L. Clinical strategies for breast cancer screening: weighing and using the evidence. Ann Intern Med 1995;122:539–47.
(*7*) Sox HC. Screening mammography in women younger than 50 years of age [editorial]. Ann Intern Med 1995;122:550–2.
(*8*) Shapiro S. The call for change in breast cancer screening guidelines [editorial]. Am J Public Health 1994;84:10–1.
(*9*) Sickles EA, Kopans DB. Deficiencies in the analysis of breast cancer screening data [editorial]. J Natl Cancer Inst 1993;85:1621–4.
(*10*) Stacey-Clear A, McCarthy KA, Hall DA, Pile-Spellman E, White G, Hulka C, et al. Breast cancer survival among women under age 50: is mammography detrimental? Lancet 1992;340:991–4.
(*11*) Curpen BN, Sickles EA, Sollitto RA, Ominsky SH, Galvin HB, Frankel SD. The comparative value of mammographic screening for women 40–49 years old versus women 50–64 years old. AJR Am J Roentgenol 1995; 164:1099–1103.
(*12*) Smart CR, Hartmann WH, Beahrs OH, Garfinkel L. Insights into breast cancer screening of younger women. Evidence from the 14-year follow-up of the Breast Cancer Detection Demonstration Project. Cancer 1993;72(4 Suppl):1449–56.
(*13*) Kopans DB. Efficacy of screening mammography for women in their forties [letter]. J Natl Cancer Inst 1994;86:1721–2.
(*14*) Peer PG, Holland R, Hendriks JH, Mravunac M, Verbeek AL. Age-specific effectiveness of the Nijmegen population-based breast cancer-screening program: assessment of early indicators of screening effectiveness. J Natl Cancer Inst 1994;86:436–41.
(*15*) Bulpitt CJ. Meta-analysis. Lancet 1988;2:93–94.
(*16*) Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis. New York: Oxford University Press, 1994:15–20.
(*17*) Elwood JM, Cox B, Richardson AK. The effectiveness of breast cancer screening by mammography in younger women [published errata appear in Online J Curr Clin Trials 1993; Doc No. 34 and 1994; Doc No. 121]. Online J Curr Clin Trials 1993; Doc No. 32.
(*18*) Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. Med J Aust 1995;162:625–9.
(*19*) Smart CR, Hendrick RE, Rutledge JH III, Smith RA. Benefit of mammography screening in women ages 40 to 49 years current evidence from randomized controlled trials [published erratum appears in Cancer 1995; 76:2788]. Cancer 1995;75:1619–26.
(*20*) Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987;9:1–30.
(*21*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.
(*22*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomized trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.
(*23*) Shapiro S. Periodic screening for breast cancer: the Health Insurance Plan project and its sequelae, 1963–1986. Baltimore: Johns Hopkins University Press, 1988.
(*24*) Tabar L, Fagerberg G, Duffy SW, Day NE, Gas A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.
(*25*) Bjurstam N, Bjornel L, Duffy SW. The Gothenburg breast screening trial: Preliminary results on breast cancer mortality for women aged 39–49. National Institutes of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD). Monogr Natl Cancer Inst 1997;22:53–55.
(*26*) Frisell J, Lidbrink E. The Stockholm mammographic screening trial: risks and benefits in age group 40–49 years. National Institutes of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD). Monogr Natl Cancer Inst 1997;22:49–51.
(*27*) Miller AB. The Canadian National Breast Screening Study: update on breast cancer mortality. National Institutes of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD). Monogr Natl Cancer Inst 1997;22:37–41.
(*28*) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer 1996;68:693–9.
(*29*) Andersson I. The Malmo mammographic screening trial: update on results and a harm-benefit analysis. National Institute of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD).
(*30*) Alexander FE. The Edinburgh randomized trial of breast cancer screening. National Institutes of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD). Monogr Natl Cancer Inst 1997;22:31–35.
(*31*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. Can Med Assoc J 1992;147:1459–76.
(*32*) Kerlikowske K, Grady D, Ernster VL. Benefit of mammography screening in women ages 40 to 49 years: current evidence from randomized controlled trials. Cancer 1995;76:1679–80.
(*33*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276:33–8.
(*34*) Bird RE. Low-cost screening mammography: report on finances and review of 21,716 consecutive cases. Radiology 1989;171:87–90.
(*35*) Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases [published erratum appears in Radiology 1992;184:878]. Radiology 1992;184:39–43.
(*36*) Burhenne HJ, Burhenne LW, Goldberg F, Hislop TG, Worth AJ, Rebbeck PM, et al. Interval breast cancers in the screening mammography program of British Columbia: analysis and classification. AJR Am J Roentgenol 1994;162:1067–71.
(*37*) Robertson CL. A private breast imaging practice: medical audit of 25,788 screening and 1,077 diagnostic examinations. Radiology 1993;187:75–9.
(*38*) Sienko DG, Hahn RA, Mills EM, Yoon-DeLong V, Ciesielski CA, Williamson GD, et al. Mammography use and outcomes in a community. The Greater Lansing Area Mammography Study. Cancer 1993;71:1801–9.
(*39*) Feig SA. Strategies for improving sensitivity of screening mammography for women aged 40 to 49 years [editorial]. JAMA 1996;276:73–4.
(*40*) Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69: 349–55.
(*41*) de Koning HJ, Boer R, Warmerdam, PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials. J Natl Cancer Inst 1995;87: 1217–23.
(*42*) de Koning HJ, Boer R. Quantitative interpretation of age-specific mortality reductions from trials by microsimulation. National Institutes of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD).
(*43*) Moskowitz M. Breast cancer: age-specific growth rates and screening strategies. Radiology 1986;161:37–41.
(*44*) Brekelmans CT, Collette HJ, Colette C, Fracheboud J, de Warrd F. Breast cancer after a negative screen: follow-up of women participating in the DOM screening programme. Eur J Cancer 1992;28A:893–5.
(*45*) Peer PG, Verbeek AL, Mravunac M, Hendriks JH, Holland R. Prognosis of younger and older patients with early breast cancer. Br J Cancer 1996;73: 382–5.
(*46*) Rajkumar SV, Sampathkumar P, Gustafson AB. Number needed to treat is a simple measure of treatment efficacy for clinicians. J Gen Intern Med 1996;11:357–9.
(*47*) Salzmann P, Kerlikowske K, Phillips K. Cost-effectiveness of extending

screening mammography programs to include women 40–49 years old. J Gen Intern Med 1997;12:63.
(48) Kopans DB. Mammography screening and the controversy concerning women aged 40 to 49. Radiol Clin North Am 1995;33:1273–90.
(49) Mettlin C, Smart CR. Breast cancer detection guidelines for women aged 40 to 49 years: rationale for the American Cancer Society reaffirmation of recommendations. CA Cancer J Clin 1994;44:248–55.
(50) American College of Radiology. Policy Statement: Guidelines for Mammography. Reston (VA): American College of Radiology, 1982.
(51) Kerlikowske K, Barclay J. Outcomes of modern screening mammography. National Institutes of Health Consensus Development Conference: Breast cancer screening for women ages 40–49. 1997 January 21–27; Bethesda (MD). Monogr Natl Cancer Inst 1997;22:105–111.
(52) Eddy DM, editor. Common screening tests. Philadelphia: American College of Physicians, 1991.
(53) U.S. Preventive Serviees Task Force. Guide to clinical preventive services (2nd ed). Baltimore: Williams & Wilkins, 1996.
(54) Canadian Task Force on the Periodic Health Examination. The periodic health examination: 2. 1985 update. Can Med Assoc J 1986;134:724–27.
(55) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.
(56) Lerman C, Tock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med 1991;114:657–61.
(57) Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996; 275:913–8.
(58) Gram IT, Lund E, Slenker SE. Quality of life following a false positive mammogram. Br J Cancer 1990;62;1018–22.
(59) Pauker SG, Kassirer JP. Contentious screening decisions: does the choice matter? [editorial]. N Engl J Med 1997336:1243–4.

Note

This work was supported by an NCI-funded Breast Cancer SPORE grant, P50 CA58207 and NCI-funded Breast Cancer Surveillance Consortium co-operative agreement, 1 U01 CA 63740.

Benefit of Screening Mammography in Women Aged 40–49: A New Meta-Analysis of Randomized Controlled Trials

*R. Edward Hendrick, Robert A. Smith, James H. Rutledge III, Charles R. Smart**

Eight randomized controlled trials (RCTs) of screening mammography have been conducted involving women aged 40–49 at entry. Current data are now available from these trials at 10.5 to 18 years of follow-up (average follow-up time: 12.7 years). Meta-analysis has been performed using a Mantel-Haenszel estimator method to combine current follow-up data from the eight RCTs of mammography that included women aged 40–49 at entry, including new follow-up data presented at the NIH Consensus Development Conference held January 21–23, 1997. Combining the most recent follow-up data on women aged 40–49 at entry into all eight RCTs yields a statistically significant 18% mortality reduction among women invited to screening mammography (relative risk: 0.82; 95% confidence interval: 0.71–0.95). Combining all current follow-up data on women aged 40–49 at entry into the five Swedish RCTs yields a statistically significant 29% mortality reduction among women invited to screening (relative risk: 0.71; 95% confidence interval: 0.57–0.89). Meta-analysis including the most recent follow-up data from all eight RCTs involving women aged 40–49 at entry demonstrates for the first time a statistically significant mortality reduction due to regular screening mammography in women of this age group. [Monogr Natl Cancer Inst 1997;22:87–92]

At the National Institutes of Health (NIH) Consensus Development Conference on Breast Cancer Screening for Women Ages 40–49, new longer-term follow-up data were presented from seven of the eight randomized controlled trials (RCTs) involving screening mammography in women aged 40–49 years at entry (*1–7*). These data updated previous results presented at the Falun Meeting in Sweden in March 1996 (*8*). All trials presented additional years of follow-up on women aged 40–49 except the Health Insurance Plan of New York (HIP) trial, which had previously published 18-year follow-up data on women 40–49 at entry (*9,10*). All trials now have follow-up data on women aged 40–49 with at least 10.5 years average follow-up since randomization.

Table 1 lists the updated subgroup data from each RCT relevant to screening mammography in women aged 40–49, the screening regimen, the number of women in the 40–49 subgroup who were entered into each arm of the trial, and the most recently presented relative risks and 95% confidence intervals from each trial. Two Swedish trials, Gothenburg and Malmö, demonstrate for the first time a statistically significant benefit from screening mammography for women under age 50 at entry. The Gothenburg trial demonstrates a statistically significant 44% mortality reduction among women 39–49 invited to screening mammography (*1*). The Malmö trial shows a statistically significant 36% mortality reduction among women aged 45–49 invited to screening mammography (*2*). Of these eight RCTs, only the Canadian National Breast Screening Study (CNBSS-1) was specifically designed to study women 40–49 at entry (*11*), and that trial now shows a slight mortality increase among women 40–49 invited to screening mammography plus clinical breast exam (*7,8*).

A previous meta-analysis of RCTs involving women 40–49, published in 1995 (*12,13*), included follow-up data ranging from 7 to 18 years since randomization (weighted average follow-up time: 10.4 years). That meta-analysis yielded a 16% mortality reduction, statistically nonsignificant at the 95% confidence level, among women 40–49 invited to screening when all eight RCTs were combined. A 24% mortality reduction, statistically significant at the 95% confidence level, was found among women aged 40–49 when all seven population-based trials were combined.

Just as the statistical power of an individual RCT increases with more participants and longer-term follow-up, the statistical power of a meta-analysis combining different trials also increases due to longer-term follow-up of individual trials. This point was noted in the Fletcher report (*14*), which acknowledged the limitations of available studies and summary analyses, stating:

> A second meta-analysis of the data from all available trials of screening in women aged 40–49 may be useful, especially when longer follow-up is available and when the effect of reclassification is clarified in the combined Swedish studies. Such a meta-analysis should use the raw data from each of the trials.

This paper presents a new meta-analysis that includes the latest follow-up data from each RCT of screening mammography in-

**Affiliations of authors:* R. E. Hendrick, Department of Radiology, University of Colorado Health Sciences Center, Denver, Colorado; R. A. Smith, Cancer Control Department, American Cancer Society, Atlanta, Georgia; J. H. Rutledge III, Department of Mathematical Sciences, United States Air Force Academy, Colorado; C. R. Smart, Salt Lake City, Utah.

Correspondence to: R. Edward Hendrick, Ph.D., Department of Radiology, C278, University of Colorado Health Sciences Center, 4200 East Ninth Avenue, Denver, Colorado 80262.

See "Note" following "References."

Table 1. Summary of RCT results for women 40–49

Study (Dates)	Screening Regimen	Frequency No. Rounds	Yrs F/U	Number of women Invited	Number of women Control	RR 95% CI
HIP Study[9] (1963–69)	2 V MM + CBE	Annually 4 rounds	18	14,432	14,701	0.77 0.53–1.11
Edinburgh[6] (1979–88)	1 or 2 V MM	24 mos 4 rounds	12.6	11,755*	10,641*	0.81* 0.54–1.20
Kopparberg[5] (1977–85)	1 V MM	24 mos 4 rounds	15.2	9,650	5,009	0.67 0.37–1.22
Östergötland[5] (1977–85)	1 V MM	24 mos 4 rounds	14.2	10,240	10,411	1.02 0.59–1.77
Malmö[2] (1976–90)	1 or 2 V MM	18–24 mos 5 rounds	12.7	13,528**	12,242**	0.64** 0.45–0.89
Stockholm[4] (1981–85)	1 V MM	28 mos 2 rounds	11.4	14,185	7,985	1.01 0.51–2.02
Gothenburg[1] (1982–88)	2 V MM	18 mos 5 rounds	12	11,724†	14,217†	0.56† 0.32–0.98
CNBSS-1[7] (1980–87)	2 V MM + CBE	12 mos 4–5 rounds	10.5	25,214	25,216	1.14 0.83–1.56

1 V MM = one-view mammography of each breast; 2 V MM = two-view mammography of each breast; CBE = clinical breast exam.

*The Edinburgh trial included three separate groups of women 45–49 at entry: the first had 5,949 women in the invited group and 5,818 in the control group (with 14 years' follow-up); the next had 2,545 in the invited group and 2,482 in the control group (12 years' follow-up); and the third had 3,261 in the invited group and 2,341 in the control group (10 years' follow-up) (*6*). Only the first group's results had been reported previously (*8*).

**The Malmö trial included two groups of women aged 45–49 at entry: one group (MMST-I) received first-round screening in 1977–8 and had 3,954 women in the invited group, 4,030 women in the control group; the second group (MMST-II) received first-round screening from 1978–90 and had 9,574 women in the invited group, 8,212 women in the control group (*2*). Only the first group's results had been reported previously (*5,8*).

†The Gothenburg trial includes women aged 39–49 at entry (*1*).

volving women aged 40–49 to assess the benefit of screening mammography in women of this age group.

Methods

A new meta-analysis of current RCT data for women aged 40–49 at entry has been performed using a Mantel-Haenszel estimator method to combine data from different trials (*15*). The Mantel-Haenszel estimator method approximates the maximum likelihood method of data pooling, with the added advantage of computational ease (*16*). The input data used for this meta-analysis are the numbers of deaths from breast cancer in both invited and control groups in each trial and the numbers of women-years of follow-up in each arm of each trial. Table 2 lists input data to the RCT meta-analysis and the references from which the most recent follow-up data were taken. The Mantel-Haenszel method weighs each trial according to the number of deaths occurring in both the invited and control groups in that trial; the greater the number of deaths, the greater weight a trial has relative to other trials included in the meta-analysis. Determinations of relative risks and confidence intervals using the Mantel-Haenszel estimator method have been based on the formalism of Breslow and Day (*17*).

Table 2. Data used in the current meta-analysis of women 40–49

Screening study	Number of Women-Years (in 1,000s) Invited	Number of Women-Years (in 1,000s) Control	Number of Breast Cancer Deaths Invited	Number of Breast Cancer Deaths Control
HIP study[9]	248	253	49	65
Edinburgh[6]	146*	135*	46*	52*
Kopparberg[5]	144	75	23	18
Östergötland[5]	143	147	27	27
Malmö[2]	166*	144*	57*	78*
Stockholm[4]	174	88	24	12
Gothenburg[1]	138†	168†	18†	39†
CNBSS-1[7]	283	283	82	72

*Included only women aged 45–49 at entry.
†Included women aged 39–49 at entry.

In cases where multiple follow-up data were available from the same trial, the data with the longest follow-up were selected for inclusion in this meta-analysis. This was determined by selecting the follow-up data that had the greatest number of breast cancer deaths among women in the invited and control groups combined.

Meta-analysis of current RCT data on mammography in women aged 40–49 were conducted under two different conditions:

1) inclusion of the most current follow-up data from all eight RCTs of mammography in women aged 40–49 at entry, and
2) inclusion of the most current follow-up data from the five Swedish RCTs of mammography in women aged 40–49 at entry.

The second meta-analysis included the five Swedish trials, each of which excluded clinical breast exam as part of its trial design (*1–5*). The HIP, Edinburgh, and CNBSS-1 trials included clinical breast exam as part of their study interventions (*9,10,6,7*), and the CNBSS-1 trial provided a clinical breast exam to all trial participants prior to their randomization into study or control groups (*11*). The five Swedish trials studied the effect of mammography alone, without the confounding influence of clinical breast examinations.

Results of our meta-analysis are stated in terms of summary relative risks (the mortality rate among women in the invited group divided by the mortality rate among women in the control group) and 95% confidence intervals (a range capturing the point estimate of relative risk 95 times if the trial or collective set of trials were repeated 100 times) determined from the combined data; 99% confidence intervals are also determined. Two-sided confidence intervals are used in each case.

Heterogeneity tests were used to assess the statistical signifi-

cance of differences among individual RCT results. The null hypothesis was that data included in the meta-analysis are homogeneous and therefore can be combined by meta-analysis without correction. A correction to the Mantel-Haenszel estimate of confidence interval is necessary if there is statistically significant evidence to reject the null hypothesis (that is, if the data are significantly heterogeneous). A chi-square test was used to assess the statistical significance of heterogeneity of individual RCT results. Breslow's random effects model was used to study the effects of possible differences among studies (*18*). The model allows for variation among studies over and above Poisson sampling errors, but without attribution to any particular factor (such as cluster randomization, screening interval, inclusion of clinical breast examination, etc.).

Results

Average follow-up time among all eight RCTs, weighted by the number of women aged 40–49 at entry in each trial, is 12.7 years. Combining the most recent follow-up data from all eight RCTs for women 40–49 years of age at entry yields the following relative risk (RR) and 95% confidence interval (95% CI):

$$\text{RR (95\% CI)} = 0.82\ (0.71\text{–}0.95).$$

This overall 18% mortality reduction among women invited to screening mammography is statistically significant at the 95% confidence level and just achieves statistical significance at the 99% confidence level (99% CI: 0.673–0.999).

Combining the most recent follow-up data from the five Swedish RCTs of women aged 40–49 at entry yields the following RR and 95% CI:

$$\text{RR (95\% CI)} = 0.71\ (0.57\text{–}0.89).$$

This 29% mortality reduction among women invited to screening mammography without clinical breast exam is also statistically significant at both the 95% and 99% confidence levels (99% CI: 0.53–0.96).

Figure 1 summarizes individual RCT results and our meta-analysis results. Bars about each relative risk point estimate in the figure represent 95% confidence intervals for individual trials and, about the two bottom points, 95% confidence intervals for the RCTs combined by meta-analysis.

Tests for statistical significance of heterogeneity of the combined RCT data demonstrate that heterogeneity is not significant among either all RCTs or the five Swedish RCTs. Chi-square tests for the heterogeneity of all eight RCTs gave $P = 0.20$; tests for the heterogeneity of the five Swedish RCTs gave $P = 0.40$. These nonsignificant results support the combination of individual RCT data by meta-analysis using the Mantel-Haenszel estimator method without correction (widening) of the 95% confidence intervals. Differences in study designs and protocols have raised the question of the effect of heterogeneity, despite the absence of statistically significant differences among RCT results. Breslow's random effects model including all eight RCTs combined yielded a relative risk of 0.81 and a 95% CI of 0.68–0.98, a slightly wider 95% CI than was given by the fixed effects model reported above. Breslow's random effects model yielded exactly the same results as the fixed effects model when the five Swedish trials were combined. These results indicate that study heterogeneity and design differences do not alter the finding of a statistically significant benefit when combining all eight RCTs involving women aged 40–49 at entry.

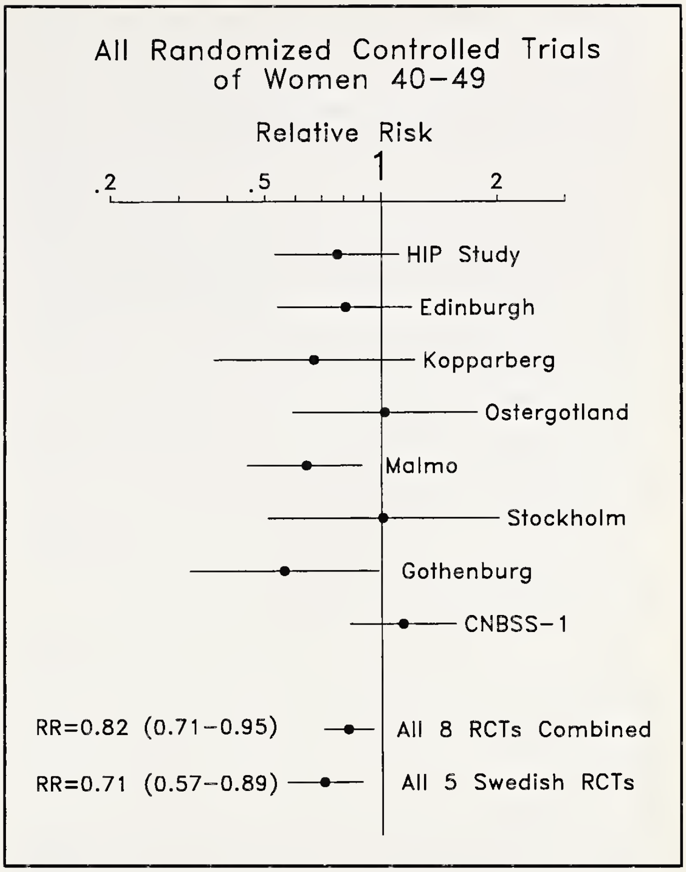

Fig. 1. Relative risks and 95% confidence intervals of all RCTs of screening mammography that included women ages 40–49 at entry. The last two data points show relative risk and 95% confidence interval results of the current meta-analysis for women ages 40–49 at entry from all eight RCTs and from the five Swedish RCTs of screening mammography.

Discussion

Current follow-up data from the eight RCTs that included women aged 40–49 at entry demonstrate delayed but increasing benefit from mammography screening. Figure 1 illustrates that two individual RCTs, the Gothenburg and Malmö trials, each have demonstrated a statistically significant mortality reduction from mammography screening among women under age 50 at entry. The Gothenburg trial included women ages 39–49 at entry (*1*), and the Malmö trial included women ages 45–49 at entry (*2*). Three other trials (HIP, Edinburgh, and Kopparberg) suggest mortality benefit to women of this age group, but the findings are not statistically significant at the 95% confidence level (*3,5,6,9,10*), and three trials (Östergötland, Stockholm, and CNBSS-1) show no benefit from screening mammography among women 40–49 (*3–5,7,8*).

It is worth examining what the entire current world's RCT data, taken collectively, say about the benefit of the invitation to screening mammography in women aged 40–49 at entry. This meta-analysis answers that question, demonstrating that all eight RCTs collectively yield a statistically significant 18% mortality

reduction among women aged 40–49 invited to screening mammography.

The major changes in individual RCT data that led to this collective demonstration of a statistically significant benefit are changes in the Gothenburg and Malmö trial results. Among women under 50, the Gothenburg trial showed a nonsignificant 27% mortality reduction at seven years' follow-up (*19,20*), a nearly significant 38% mortality reduction at 10 years' follow-up (*8*), and a statistically significant 44% mortality reduction at 12 years' follow-up (*1*). The most recent data reported by Malmö investigators have included results from the so-called MMST-II group, an additional 17,000 women randomized at ages 45–48 and entered into the study between 1978 and 1990 (*2*). These additional 17,000 women, added to the approximately 7,000 women ages 45–49 randomized and reported on previously (the MMST-I group) (*5*), have significantly boosted the statistical power of the Malmö trial results (*2*), producing a statistically signifieant 36% mortality reduction from the combined Malmö (MMST-I and MMST-II) trial results.

Results for the subgroup of women aged 40–49 at entry from the HIP trial (*9,10*), the Edinburgh trial (*6*), and the combined Swedish trials (*20,21*) indicate that as more years of follow-up are included, benefit eventually emerges and there is a steady progression toward greater benefit from screening mammography. Meta-analyses of the eight RCTs show this same trend. Cox's meta-analysis of RCT data on women 40–49 at approximately seven years of follow-up showed no benefit, yielding a relative risk of 1.04 (95% CI: 0.81–1.33) when all eight RCTs were combined (*22*). At seven to nine years of follow-up, Kerlikowske's meta-analysis of RCT data on women 40–49 gave a similar relative risk of 1.02 (95% CI: 0.82–1.27) (*23*). Our previous meta-analysis of RCT data on women 40–49, at an average of 10.4 years follow-up, gave a 16% mortality reduction from the invitation to screening to women 40–49 in all eight RCTs (*12,13*). The current meta-analysis, at an average of 12.7 years of follow-up, gives an 18% mortality reduction from invitation to screening to women 40–49 from all eight RCTs, statistically significant at the 95% confidence level for the first time.

It has been pointed out previously that the potential benefit of screening mammography takes longer to manifest in women aged 40–49 than in older women (*12,20*). A delayed demonstration of benefit is to be expected in women 40–49 years of age compared to older women due to fewer breast cancer deaths for the following reasons:

1) breast cancer incidence and mortality rates are lower in women 40–49 than in women 50 and over;
2) the number of women 40–49 included in the eight RCTs is approximately one-third the total number of women included in the eight trials;
3) the higher rates of ductal carcinoma *in situ* (DCIS) in women 40–49 than in older women and the slow progression of DCIS to invasive carcinoma require a longer time to manifest a mortality difference between screen-detected DCIS in the study group and undetected DCIS in the control group.

A delayed demonstration of benefit in women 40–49 is also to be expected due to somewhat less favorable cancer stage distributions resulting from use of a wide screening interval in some RCTs:

4) on average, the lead time of mammography is shorter in women 40–49 than in women 50 and over;
5) the sensitivity of mammography in the RCTs is known to be lower in women 40–49 than in women 50 and over (*14*);
6) a longer period of follow-up will be needed if the benefit from screening mammography in the trials among women 40–49 was limited to cancers detected with good to intermediate prognosis. Recent analyses of the Swedish two-county data have shown that the two-year screening interval used in these two trials (Kopparberg and Östergötland) was not effective in detecting more aggressive tumors with poor prognosis (*8*). These findings, in conjunction with previous analyses estimating age-specific mean sojourn times, support the conclusion that annual screening is necessary to achieve mortality reductions in women 40–49 similar to those obtained in women 50 and over with wider screening intervals (*24*).

These factors influence the outcomes of trials for women 40–49 and make it more difficult to demonstrate a statistically significant mortality reduction in them as compared with women 50 and older. Henee, longer follow-up is needed to manifest a statistically significant mortality reduetion in women aged 40–49.

Because of the delayed benefit of screening mammography in women 40–49, some have argued that the observed benefit of mammography among women 40–49 at randomization may be due to "age migration": the effect that women 40–49 at entry may benefit in terms of mortality reduction only from screening mammography performed at or after the age of 50 (*25,26*). Age migration is an inevitable consequence of randomizing a wider age range of women, screening them over a number of years, and then attempting to perform subgroup analyses of trial results based on age at entry. While it may be interesting to examine trial data in terms of age at diagnosis rather than age at entry, it is methodologically unsound to do so. As Prorok *et al.* point out, age at diagnosis is a pseudovariable, since it is influenced by the study intervention (screening) during the trial, reducing the comparability of the study and control groups (*27*). Thus, however intriguing, it is not clear that any results from subgroup analyses based on age at diagnosis are credible. Moreover, such analyses only further subdivide original data sets that have already been subdivided by age at entry, completely eliminating any possible statistieal power of the data. Nevertheless, data in the published literature and presented at the NIH Consensus Conference do not support the age migration hypothesis that benefit among women 40–49 at entry is due to the subset of women diagnosed after age 50. Tabar *et al.* compared invited and control groups from the Swedish two-county trial based on age at diagnosis and showed a 15% mortality benefit among women both randomized and diagnosed before age 50, compared with only a 5% benefit among women randomized in their forties and diagnosed in their fifties (*28*). The mortality difference is actually higher among women diagnosed in their forties than among women diagnosed in their fifties in the Swedish two-county trial.

The suggestion that much of the benefit to women invited to screening within RCTs results from clinical breast exams that were included along with screening mammography is also spe-

cious. None of the five Swedish trials included clinical breast exams, yet previously combined results of those five trials demonstrated a 23% mortality reduction from screening, just barely lacking statistical significance at the 95% confidence level: RR (95% CI) = 0.77 (0.59–1.01) (*5,8*). Including all new follow-up data presented at the NIH Consensus Conference, combined data from the five Swedish trials yields a 29% mortality reduction for women under 50 at entry, statistically significant at the 95% confidence level: RR (95% CI) = 0.71 (0.57–0.89). These results indicate that clinical breast exams play an insignificant role in the mortality reductions observed in RCTs.

The true benefit of mammography today is likely to exceed the benefit demonstrated in RCTs for at least two reasons:

1) RCTs test the efficacy of the invitation to screening mammography in a predefined study group compared to no invitation in a predefined control group. In population-based RCTs that measured compliance among women offered screening, compliance rates for the first screening mammogram ranged from 61% to 89%, with lower compliance rates in each subsequent screen. Since a statistically significant benefit from mammography in women 40–49 has been shown to exist, the true benefit to women receiving regular screening mammography will be greater than the benefit demonstrated among women in the RCTs invited to screening mammography, since a reasonable fraction of women invited to screening did not comply. Likewise, women who were assigned to the control group but who went outside the trial to obtain regular screening mammography diluted the observed benefit of screening in the RCTs, providing a second reason why the true benefit of regular screening mammography will be greater than the demonstrated benefit (*29*).

2) The technology of mammography has improved markedly since the time of even the most recent RCTs. Women receiving regular, high-quality mammography today are more likely to have their cancers detected at smaller sizes and at earlier stages than women who participated in the eight RCTs, as illustrated by comparing the surrogate prognostic indicators of mammography as practiced today in the United States to those same indicators in any of the eight RCTs. Sickles (*30*) and Linver (*31*) have presented prognostic indicators of modern mammography in clinical practice in women 40–49, comparing them to the results of RCTs, suggesting that modern mammography in the United States should do a better job of detecting cancers and saving lives in women 40–49 than did the RCTs.

Conclusions

With the latest follow-up data from RCTs involving women 40–49, there is now convincing evidence of benefit from screening mammography to women of this age group. A statistically significant mortality reduction is shown at the 95% confidence level for women 40–49 at entry from two of the eight individual RCTs (Gothenburg and Malmö), from the combined data on women 40–49 from all eight RCTs, and from the combined data on women 40–49 from the five Swedish trials. These results indicate that screening mammography was effective in reducing breast cancer deaths among women 40–49 at entry with or without clinical breast exams, even with noncompliance of some women in the invited groups and mammography outside the trials among some women in the control groups. Even greater benefits should accrue today from regular screening mammography in women ages 40–49 than has been demonstrated by the collective results of the eight randomized controlled trials.

References

(*1*) Bjurstam N, Bjorneld L, Duffy SW. The Gothenburg Breast Screening Trial: results from 11 years follow-up. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:63–4.

(*2*) Andersson I. The Malmö Mammographic Screening Trial: update on results and a harm-benefit analysis. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:51–3, and private communication.

(*3*) Tabar L, Fagerberg G, Duffy SW. Recent results from the Swedish Two-County Trial: the effects of age, histologic type, and mode of detection. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:55–7.

(*4*) Frisell J, Lidbrink E. The Stockholm Mammographic Screening Trial: risks and benefits. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:59–61.

(*5*) Nystrom L, Wall S, Rutqvist L, Andersson I, Bjurstam N, Fagerberg G, et al. Update of the overview of the Swedish Randomized Trials on breast cancer screening with mammography. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:65–9.

(*6*) Alexander FE. The Edinburgh Randomized Trials of breast cancer screening. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:49, and private communication.

(*7*) Miller AB. The Canadian National Breast Screening Study: update on breast cancer mortality. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:51–3, and private communication.

(*8*) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer 1996;68:693–9.

(*9*) Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer: The Health Insurance Plan project and its sequelae; 1963–1986. Baltimore (MD): Johns Hopkins University Press, 1988.

(*10*) Shapiro S. Periodic Screening for Breast Cancer: The Health Insurance Plan of Greater New York Randomized Controlled Trial. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:41–8.

(*11*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years [published erratum appears in Can Med Assoc J 1993;148:718]. Can Med Assoc J 1992;147:1459–76.

(*12*) Smart CR, Hendrick RE, Rutledge JH III, Smith RA. Benefit of mammography screening in women ages 40 to 49 years. Current evidence from randomized controlled trials [published erratum appears in Cancer 1995; 75:2788]. Cancer 1995;75:1619–26.

(*13*) Smart CR, Hendrick RE, Rutledge III JH, Smith RA. Benefit of mammography screening in women ages 40–49 years: Current evidence from randomized controlled trials. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:83–9.

(*14*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.

(*15*) Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.

(*16*) Rothman KJ. Modern epidemiology. Boston: Little, Brown, & Co., 1986: 195–236.

(*17*) Breslow NE, Day NE. Statistical methods in cancer research: Volume II—The design and analysis of cohort studies. Oxford: Oxford University Press, 1987:109–113.

(*18*) Breslow NE. Extra-Poisson variation in log-linear models. Applied Statistics 1984;33:38–44.

(*19*) Wald N, Chamberlain F, Hackshaw A. Report of the European Society for

Mastology: Breast Cancer Screening Evaluation Committee (1993). Breast 1993;2:209–16.
(20) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.
(21) Nystrom L, Larsson LG. Breast cancer screening with mammography [letter]. Lancet 1993;341:1531–2.
(22) Cox B. Variation in the effect of breast cancer screening by year of follow-up. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:71–3.
(23) Kerlikowske KM. Efficacy of screening mammography: relative and absolute benefit. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:77–81.
(24) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.
(25) de Koning HJ, Boer R, Warmerdam PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials. J Natl Cancer Inst 1995;87:1217–23.
(26) de Koning HJ. Quantitative interpretation of age-specific mortality reductions from trials by microsimulation. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:93–96.
(27) Prorok PC, Hankey BF, Bundy BN. Concepts and problems in the evaluation of screening programs. J Chronic Dis 1981;34:159–71.
(28) Tabar L, Duffy SW, Chen HH. Re: Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer screening trials. J Natl Cancer Inst 1996;88:52–5.
(29) Glasziou PP. Meta-analysis adjusting for compliance: the example of screening for breast cancer. J Clin Epidemiol 1992;45:1251–6.
(30) Sickles EA. Screening outcomes: clinical experience with service screening using modern mammography. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:105–110.
(31) Linver MN. Mammography outcomes in a practice setting by age: prognostic factors, sensitivity, and positive biopsy rate. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:115–9.

Note

We would like to thank Mr. Stephen W. Duffy for valuable discussions and his insights concerning heterogeneity corrections and random effects models for analyzing data from multiple RCTs.

Markov Models of Breast Tumor Progression: Some Age-Specific Results

*Stephen W. Duffy, Nicholas E. Day, László Tabár, Hsiu-Hsi Chen, Teresa C. Smith**

Researchers have noted that mammographic screening has a reduced effect on breast cancer mortality in women in their forties compared to older women. Explanations for this include poorer sensitivity in younger women due to denser breast tissue, as well as more rapid tumor progression, giving a shorter mean sojourn time (the average duration of the preclinical screen-detectable period). To test these hypotheses, we developed a series of Markov-chain models to estimate tumor progression rates and sensitivity. Parameters were estimated using tumor data from the Swedish two-county trial of mammographic screening for breast cancer. The mean sojourn time was shorter in women aged 40–49 compared to women aged 50–59 and 60–69 (2.44, 3.70, and 4.17 years, respectively). Sensitivity was lower in the 40–49 age group compared to the two older groups (83%, 100%, and 100%, respectively). Thus, both rapid progression and poorer sensitivity are associated with the 40–49 age group. We also modeled tumor size, node status, and malignancy grade together with subsequent breast cancer mortality and found that, to achieve a reduction in mortality commensurate with that in women over 50, the interscreening interval for women in their forties should be less than two years. We conclude that Markov models and the use of tumor size, node status, and malignancy grade as surrogates for mortality can be useful in design and analysis of future studies of breast cancer screening. [Monogr Natl Cancer Inst 1997:22: 93–97]

In assessing the early detection of a disease through screening, a first model is often the following:

(1) Every subject begins with no detectable disease at all. Some subjects will develop the disease of interest, some will remain free of the disease all their lives.
(2) For a subject who develops disease, at a certain time t_1, the person will pass to a state in which the disease is asymptomatic but can be detected by a screening test. This phase is often called the preclinical detectable period (PCDP).
(3) For this subject, at a certain time t_2 ($t_2 > t_1$), the disease will become clinically symptomatic. In the absence of screening, this is defined as the time of diagnosis (although in practice there may be a delay from symptoms to diagnosis). The period t_2-t_1 is known as the sojourn time.

Screening might take place as part of an immunization program, to prevent the spread of a communicable disease or to identify cases in time to effectively treat them. Here we will concentrate on the last purpose, and the specific area we will focus on is breast cancer screening with mammography. For screening to be effective in this context, disease needs to be diagnosed some time before t_2, while it is still treatable with less aggressive methods and while it is curable in the long term. This means a substantial lead time and good sensitivity are required. Lead time = $t_2 - t_3$, where t_2 is time of clinical diagnosis as above and t_3 is actual time of detection by screening. Sensitivity is the probability that a case of preclinical, detectable disease is actually diagnosed by the screening test. The sensitivity and the average length of the preclinical detectable period (= mean sojourn time [MST]) are therefore crucial parameters in assessing the ability of screening to affect subsequent mortality.

Note that the sojourn time is an upper limit on the lead time achievable, but if sojourn time is assumed to be exponentially distributed, the expected lead time of a screen-diagnosed cancer is equal to the mean sojourn time. The seminal papers on this subject are by Zelen and Feinlieb (*1*), Prorok (*2*), and Day and Walter (*3*).

Usually, in the modeling of tumor progression and its arrest by early detection, estimation has to be heavily supported by assumptions, constraints, and analytic strategies one would prefer to avoid. These include estimation in several stages—for example, the underlying preclinical incidence may be estimated as the clinical incidence in an unscreened population and the progression rate to clinical disease estimated thereafter (*3*), or the rate of progression to clinical disease may be estimated first assuming a 100% sensitivity of the early-detection tool and the sensitivity thereafter estimated with the progression rate assumed constant at the estimated value (*4*). It has also often been necessary to make sweeping assumptions about sensitivity (*5*). Loss of information due to blocking of time into discrete years or screening rounds is also common (*3,6*). One well-designed method employed in the past is that of Day and Walter (*3*), which estimates the sensitivity and progression rate simultaneously but which requires a prior estimate of the underlying disease incidence to be specified as fixed beforehand.

It seems intuitively desirable to develop a comprehensive model that would simultaneously estimate all parameters, if a data set of adequate design could be found. In particular, it would be desirable to estimate the preclinical incidence from the

**Affiliations of authors:* S. W. Duffy, N. E. Day, T. C. Smith, MRC Biostatistics Unit, Institute of Public Health, Cambridge, United Kingdom; L. Tabár, Director, Mammography Department, Central Hospital, Falun, Sweden; H. H. Chen, Taiwan National University, Taipei, Taiwan.

Correspondence to: Stephen W. Duffy, M.D., MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK.

same data set used to estimate the progression from the preclinical state. Here we demonstrate the use of Markov-chain models to estimate the progression rates from empirical screening data and point up some applications in assessing the likely age-specific effect of screening on future mortality from breast cancer.

Data and Methods

We used the data from the Swedish two-county trial of mammographic screening for breast cancer (*7*); 77,080 women aged 40–74 were randomized to invitation to screening (Active Study Population [ASP]), and 55,985 to no invitation (Passive Study Population [PSP]), for seven to eight years. The PSP was give a single screen at the last screen of the ASP. We shall concentrate on women aged 40–69 at randomization, as screening was abandoned after the second screen in women aged 70–74 due to poor attendance rates. The cancers diagnosed in the trial are shown by detection mode in Table 1.

Progression of the disease was modeled as a Markov chain (*8*). In this model, individuals occupy states for random, exponentially distributed periods of time and move from state to state independently of each other. The major assumption of this model is that if we know the state at time t for a given individual, knowledge of that individual's states at times prior to t is of no additional benefit in assessing the individual's likely future progression.

A simple example is a three-state model where states 0, 1, and 2 represent no detectable disease, preclinical screen-detectable disease, and clinical symptomatic disease, respectively. Associated with such a Markov model is a transition matrix of instantaneous probabilities of moving from state to state. For the above three-state model we posit the following transition matrix:

$$\begin{bmatrix} -\lambda_1 & \lambda_1 & 0 \\ 0 & -\lambda_2 & \lambda_2 \\ 0 & 0 & 0 \end{bmatrix}$$

Here λ_1 denotes the birth rate into the PCDP and λ_2 the transition rate from preclinical to clinical disease. We assume spontaneous regression to be impossible. We also assume that to reach the clinical phase, a tumor must pass through the preclinical phase. A property of this model is that $1/\lambda_2$ is the MST (*9*). The instantaneous transition rates need to be converted into probabilities of transition in noninstantaneous periods of time, by solving a potentially complex set of algebraic equations known as Kolmogorov equations (*8*). In this simple model, the solution can be derived by hand, giving the formula for probabilities of transition in a non-negligible time t as:

$$\begin{bmatrix} e^{-\lambda_1 t} & \dfrac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{(\lambda_1 - \lambda_2)} & 1 - \dfrac{\lambda_1 e^{-\lambda_2 t} - \lambda_2 e^{-\lambda_1 t}}{(\lambda_1 - \lambda_2)} \\ 0 & e^{-\lambda_2 t} & 1 - e^{-\lambda_2 t} \\ 0 & 0 & 1 \end{bmatrix}$$

The formulas will become further complicated by the fact that those with a previous history of clinical breast cancer were excluded, so we have to condition on this at the first screen of the ASP and at entry to the trial of the PSP. Also, inclusion of false-positive and false-negative screening error probabilities render the probabilities very complex indeed.

Also, introducing more states—for example, "node negative" and "node positive" within each of the preclinical and clinical phases—brings about considerable increases in algebraic complexity. In the latter case, hand solution of the Kolmogorov equations is not feasible, so the following strategy was implemented:

(1) We used the computer program *Mathematica* to solve the Kolmogorov equations to produce transition probabilities from the transition rates (*10*).
(2) For each type of transition observed, we used (1) and the error probabilities to calculate the expected number of transitions.
(3) We then solved, as a nonlinear regression, the equation: observed transitions = expected transitions + error.

This means we did not actually maximize the likelihood but instead estimated the transition rates and error probabilities as a solution of a complex set of generalized estimating equations. For further details of algebra and statistical methods, see Duffy et al. (*9*) and Chen et al. (*11,12*).

Results

Three-State Model

Table 2 shows the results for a model of progression among three states, as described above: no detectable disease, preclini-

Table 1. Cancers in the two-county study by detection mode (%) and age

Detection mode	Age 40–49	50–59	60–69	70–74
ASP prior*	6 (2)	5 (1)	13 (2)	4 (1)
ASP screen 1	39 (15)	103 (27)	184 (35)	101 (39)
ASP screen 2+	110 (43)	156 (41)	183 (35)	52 (20)
ASP interval	91 (36)	90 (24)	96 (18)	52 (20)†
ASP refuser	10 (4)	28 (7)	53 (10)	50 (20)
Total ASP	256	382	529	259
PSP pre-screen	115 (71)	221 (71)	277 (66)	142 (96)
PSP screen	47 (29)	94 (29)	140 (34)	6 (4)
Total PSP	162	315	417	148

*Prior = cancers diagnosed clinically between randomization and first screen.
†Includes 30 cancers diagnosed after screening was abandoned in this age group.

Table 2. Three state model results—instantaneous transition rates, MST, senstivity, and PPV

Parameter	Age 40–49	50–59	60–69
Preclinical incidence rate per 100,000 person-years (95% CI)	89 (84–95)	155 (150–160)	240 (230–251)
MST in years (95% CI)	2.44 (2.12–2.86)	3.70 (3.44–4.17)	4.17 (4.00–4.55)
Sensitivity (95% CI)	83% (76–91%)	100% (–)	100% (–)
PPV	85%	100%	100%

cal screen-detectable disease, and symptomatic clinical disease. The rate of progression from preclinical to clinical disease (the reciprocal of the mean sojourn time) is much faster in the 40–49 age group than in women aged 50 or more. Note that in Table 2, we present the false-positive probability in terms of positive predictive value (PPV)—that proportion of screen-detected tumors that would have arisen clinically in the future had screening not taken place. PPV is commonly used to evaluate diagnostic tests, and should not be confused with biopsy predictive value. In women aged 50 or more, both sensitivity and PPV were around 100%, whereas in the 40–49 group, sensitivity was 83% and PPV 85%.

Five-State Model

Consider a model including axillary lymph node status. There are five states:

(1) No detectable disease (0);
(2) Preclinical node negative (pre −);
(3) Preclinical node positive (pre +);
(4) Clinical node negative (clin −);
(5) Clinical node positive (clin +);

The transition matrix is:

$$\begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & 0 & 0 \\ 0 & -\lambda_2-\lambda_3 & \lambda_2 & \lambda_3 & 0 \\ 0 & 0 & -\lambda_4 & 0 & \lambda_4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We assume no regression, as before, that all tumors are born node negative and in the preclinical phase and that transition in two dimensions at exactly the same instant is not possible. Note that we cannot estimate transitions within the clinical phase, as once a tumor is diagnosed, it is excised and further assessment of natural history thereafter is impossible.

Table 3 shows the estimated instantaneous transition rates. Note the more rapid progression from node-negative to node-positive tumors in the preclinical phase in younger women and the more rapid progression from preclinical to clinical. In all age groups, the inclusion of more states to pass through once a cancer is in the preclinical state leads to a faster rate of progression into the preclinical state.

Table 3. Results for a five-state model for progression with respect to node status model

Transition	Rate (95% CI) for age 40–49	Rate (95% CI) for age 50–59	Rate (95% CI) for age 60–69
0 → preclinical N−	0.00122 (0.00120–0.00125)	0.00176 (0.00175–0.00177)	0.00263 (0.00260–0.00267)
preclinical N− → preclinical N+	0.35 (0.22–0.55)	0.23 (0.15–0.35)	0.15 (0.10–0.23)
preclinical N− → clinical N−	0.26 (0.16–0.42)	0.18 (0.11–0.29)	0.20 (0.13–0.30)
preclinical N+ → clinical N+	2.11 (1.09–4.08	0.85 (0.54–1.33)	0.61 (0.41–0.91)

In terms of probabilities of progression within one year, we have the following transition probability matrices:

(a) 40–49

$$\begin{bmatrix} 0.9988 & 0.0009 & 0.00009 & 0.00014 & 0.00008 \\ 0 & 0.54 & 0.10 & 0.20 & 0.16 \\ 0 & 0 & 0.12 & 0 & 0.88 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Thus, for example, a node-negative preclinical tumor has a 10% chance of becoming node positive but remaining preclinical in one year (second row, third column).

(b) 50–59

$$\begin{bmatrix} 0.9983 & 0.0014 & 0.00014 & 0.00014 & 0.00004 \\ 0 & 0.66 & 0.12 & 0.15 & 0.07 \\ 0 & 0 & 0.43 & 0 & 0.57 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(c) 60–69

$$\begin{bmatrix} 0.9974 & 0.0022 & 0.00017 & 0.00018 & 0.00005 \\ 0 & 0.71 & 0.11 & 0.13 & 0.05 \\ 0 & 0 & 0.46 & 0 & 0.54 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Implications of the three transition probability matrices include:

(1) In the age group 40–49, a tumor which is node negative and preclinical now has a 46% chance of progression to node positive or clinical phase or both within the next year. The corresponding figures for the 50–59 and 60–69 age groups are 34% and 29%, respectively.
(2) For preclinical node-positive tumors in the 40–49 age group, 88% progress to the clinical phase within a year—that is, there is very little opportunity for detection by screening thereafter. For the 50–59 and 60–69 age groups, the figures are 57% and 54%, respectively.
(3) A preclinical node-negative tumor in the 40–49 age group is about three times as likely to progress to clinical node positive than a corresponding preclinical node-negative tumor in the 50+ age groups.

Similar patterns are observed for tumor size and for a model that includes progression with respect to both variables.

Malignancy Grade

This is a histological measure of aggressive potential of the tumor comprising differentiation, nuclear size, pleomorphism, and mitotic rate. Tumors are graded as 1 (good prognosis), 2 (intermediate prognosis), or 3 (poor prognosis). The malignancy grade used to be thought of as an innate unchanging quantity, but this may be an oversimplification because of a phenomenon known as "dedifferentiation," or "phenotypic drift" (*13*). It is

well documented that some tumors are internally heterogeneous with respect to grade (and indeed phenotypic character). In this case, the pathologist has to score the grade based on what is the dominant component of the tumor examined. One might suspect that in such a case, the more aggressive component would grow faster than the less aggressive if the tumor were left untreated—that is, if the malignancy grade changes (the tumor differentiates) as the cancer ages. This would be manifested by more grade 3 tumors in a control series than a screened series, after elimination of length-bias cases. Length bias is removed by excluding the first screen from both the ASP and the PSP. This is because: first, theory tells us that the length bias cases remain in the preclinical screen-detectable phase for a long time and therefore need only one screen to detect them; second, in the two-county study, the excess incidence in the ASP vanished after a single screen of the PSP; and third, the experience of clinicians working with screening is that the first screen contains a disproportionate number of dubious malignancies.

Table 4 shows the percentage of grade 3 tumors by age and study group, after removal of the length-bias cases. There does indeed seem to be a tendency for tumors to dedifferentiate.

There may be a further complication in that some tumors have this heterogeneity and therefore the potential to "dedifferentiate" and others do not. We therefore propose a mover-stayer mixture of models (*12*). Suppose we have a five-state model:

(1) No detectable disease.
(2) Preclinical grade 1–2.
(3) Preclinical grade 3.
(4) Clinical grade 1–2.
(5) Clinical grade 3.

For an unknown proportion p of tumors, the transition rates from state (2) to (3) and from state (4) to (5) are zero (i.e., changing grade with time is impossible), and for the remaining 1-p of all tumors, nonzero transition rates apply (i.e., progression with respect to grade is possible). Fitting this model to the two-county data, we estimated the proportion of tumors with the propensity to dedifferentiate to be 81%, 48%, and 51% for the age groups 40–49, 50–59, and 60–69, respectively. Thus, there is a larger proportion of tumors whose malignancy grade may deteriorate in women aged under 50.

Discussion

The overwhelming implication of the above results is that progression to the clinical phase, and with respect to node status and tumor size (data available from the authors), is faster in the age group 40–49 than in older age groups. In addition, the potential for dedifferentiation or phenotypic drift is stronger in the 40–49 age group than in women aged 50 or more. This is consistent with previous suggestions that a shorter interscreening interval is required in this age group. It is of some value to quantify this further, in terms of the mortality expected from different screening frequencies. We used the survival data from the 2,468 tumors in the two-county study to predict mortality as follows:

(1) Using the Markov models, we predicted the numbers of tumors by node status, size, and grade in an unscreened population and in a population screened every one, two, or three years.
(2) Using survival data to estimate the Cox regression parameters for the various categories, we estimated the 10-year survival probability in each category.
(3) We then multiplied the expected numbers of tumors in each category by the proportion expected to die of breast cancer in that category, thus giving the expected 10-year mortality.
(4) Finally, we obtained the predicted relative mortality by dividing the predicted mortality for the screened population by that for the unscreened.

Table 5 shows the predicted relative mortality using a model incorporating both size and node status. The effects of annual two-year and three-year screening are given. In our calculations, we assumed that 90% of those invited actually attended for screening, and we used the sensitivities estimated in Table 2. Major points to note are that the predicted effects are close to those observed in the two-county study, that a shorter interscreening interval is required in the age group 40–49, and that the interval is less crucial for older women.

We can validate the use of the two-county study survival data by applying them to other trials to predict the relative mortality. Figure 1 shows the relative mortality observed in the Malmö, Gothenburg, Edinburgh, two-county, Stockholm, and Canadian trials, compared with that predicted using the node status, tumor size, and (where available) malignancy grade of tumors diagnosed within each trial, coupled with the survival rates pertaining to node status, size, and grade from the two-county study (*14*). The line of perfect agreement is also shown. Clearly the agreement between predicted and observed relative mortality is good, and in five out of the six trials, the predicted mortality gives a slightly conservative result.

The above has implications for study design. First, the Markov models and predicted mortality methods may be used for power and sample size calculations. Second, because of the greater information, predicted mortality from tumors diagnosed has a lower variance than observed mortality. We might therefore consider using the predicted mortality from the tumors diagnosed as a surrogate in studies to evaluate breast cancer screening strategies. Predicted mortality is to be used in the UK

Table 4. Percent of grade 3 tumors by age and study group, two-county trial

Group	40–49	50–59	60–69
Bias-free ASP	45%	38%	39%
Bias-free PSP	51%	49%	46%

Table 5. Expected relative ten-year mortalities from 3-yearly, 2-yearly, and annual screening by age group*

Interval between screens	40–49 (83% sensitivity)	50–59 (100% sensitivity)	60–69 (100% sensitivity)
1 year	0.64	0.54	0.56
2 years	0.82 (0.87)	0.61	0.61
3 years	0.96	0.66 (0.66)	0.66 (0.60)

*Relative mortality calculated as deaths/person-years for the invited group divided by the same figure for the control group, assuming sensitivities as in Table 2 and 90% attendance rates. Figures in parentheses represent the observed relative mortalities in the Swedish Two-County Study.

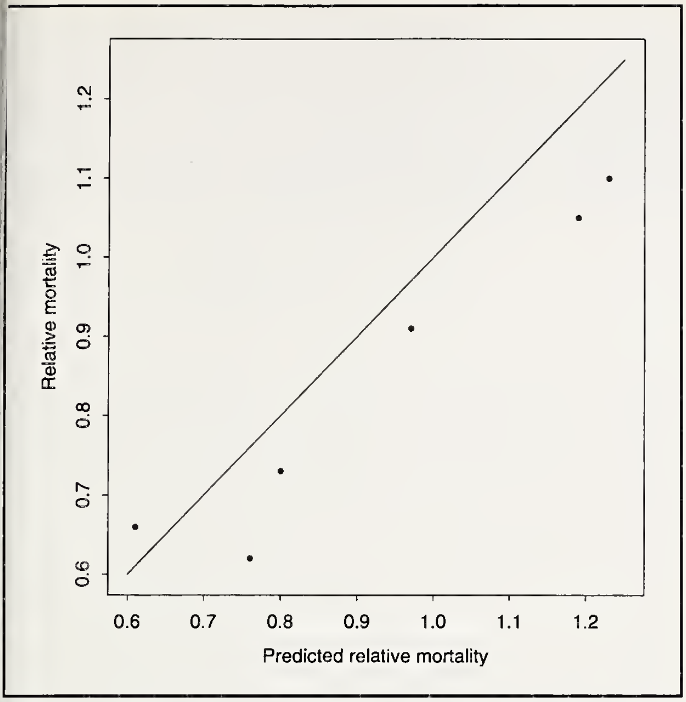

Fig. 1. Observed relative mortality from breast cancer in the 40–49 age group plotted against that predicted from the size, node status, and (where available) malignancy grade of tumors diagnosed, for the Malmö, Gothenburg, Edinburgh, Two-County, Stockholm, and Canadian trials.

Breast Screening Frequency Trial rather than actual mortality, and preliminary analysis indicates that this will double its power (*15*). Predicted mortality also provides results some 10 years earlier than observed mortality. This is particularly relevant to the case of breast cancer screening in the age group 40–49, for whom the actual mortality effect is often far off in the future, but the need for an answer is relatively urgent.

Conclusions

We draw the following conclusions from our analysis:

(1) Progression to a more advanced state is considerably more rapid in the 40–49 age group.

(2) This progression also occurs with respect to the malignancy grade of the tumor. The proportion of tumors capable of dedifferentiation appears to be greater in women aged 40–49.

(3) In this age group, the best indicator of future benefit is the relative rate of advanced tumors, or the predicted deaths from these. These can reasonably be used in trials.

(4) There is a potential for the effect on advanced tumors to be used to assess the likely future effect on mortality, but only if there are good data available on the stage, size, or node status of tumors before screening.

(5) The above results do not tell us whether or not to screen in this age group. They do, however, tell us something of the biological background that screening in this age group is up against and indicate that a shorter interscreening interval is more likely to be effective.

References

(*1*) Zelen M, Feinleib M. On the theory of screening for chronic disease. Biometrika 1969;56:601–14.

(*2*) Prorok PC. The theory of periodic screening II. Doubly bounded recurrence times and mean time and detection probability estimation. Adv Appd Probability 1976;8:460–76.

(*3*) Day NE, Walter SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. Biometrics 1984; 40:1–14.

(*4*) Paci E, Duffy SW. Modelling the analysis of breast cancer screening programmes: sensitivity, lead time and predictive value in the Florence District Programme (1975–1986). Int J Epidemiol 1991;20:852–8.

(*5*) de Koning HJ, Boer R, Warmerdam PG, Beemsterboer PM, van der Maas PJ. Quantitative interpretation of age-specific mortality reductions from the Swedish breast cancer-screening trials. J Natl Cancer Inst 1995;87: 1217–23.

(*6*) Chen JS, Prorok PC. Lead time estimation in a controlled screening program. Am J Epidemiol 1983;118:740–51.

(*7*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.

(*8*) Cox DR, Miller HD. The theory of Stochastic Processes. London: Methuen, 1965.

(*9*) Duffy SW, Chen HH, Tabar L, Day NE. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. Stat Med 1995;14:1531–43.

(*10*) Wolfram S. Mathematica: A system for doing mathematics by computer. Redwood City (CA): Addison-Wesley, 1991.

(*11*) Chen HH, Duffy SW, Tabar L. A Markov chain method to estimate the tumor rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. Statistician 1996;45:307–17.

(*12*) Chen HH, Duffy SW, Tabar L. A mover-stayer mixture of Markov chain models for the assessment of dedifferentiation and tumor progression in breast cancer. J Appd Stat (in press).

(*13*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Gad A. Tumor development, histology and grade of breast cancers: prognosis and progression. Int J Cancer 1996;66:413–9.

(*14*) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer 1996;68:693–9.

(*15*) Day NE, Duffy SW. Trial design based on surrogate endpoints: application to comparison of different breast screening frequencies. J Roy Statist Soc A 1996;159:49–60.

Breast Cancer Screening Outcomes in Women Ages 40–49: Clinical Experience With Service Screening Using Modern Mammography

*Edward A. Sickles**

The several randomized controlled trials (RCTs) of breast cancer screening among women of ages 40 to 49 now collectively show a statistically significant reduction in breast cancer mortality. However, there have been numerous recent advances in mammography, such that it now is demonstrably better than when the RCTs were conducted. The use of surrogate measures of screening efficacy (tumor size, lymph node status, cancer stage), readily derived from modern service screening programs, demonstrates how the improved mammography of the 1990s should produce a greater degree of mortality reduction among women ages 40–49 than that already demonstrated in the RCTs. Indeed, these surrogate measures of mortality reduction are as favorable for women of ages 40–49 and 65+ as they are for women of ages 50–64, strongly suggesting that, since modern service screening is accepted as effectively reducing mortality among women of ages 50–64, it should also effectively reduce mortality among women in the 40–49 and 65+ age groups. [Monogr Natl Cancer Inst 1997;22:99–104]

The best evidence of mortality reduction from breast cancer by screening women ages 40 to 49 comes from the several randomized controlled trials (RCTs) that already have been conducted. Like the screening carried out in the RCTs, service screening is performed on entire populations of women, either by invitation or by self-selection. However, service screening does not provide data from randomly selected control groups of nonscreened women. Therefore, service screening programs do not generate outcomes data that are sufficiently rigorous to independently furnish convincing evidence on mortality reduction.

Nonetheless, there still is considerable value in the outcomes data from modern service screening programs, for several reasons: (a) the data from existing RCTs indicate the presence of a substantial and (as of January 1997) statistically significant mortality reduction, but controversy remains over the magnitude of this mortality reduction; (b) known deficiencies in design and execution of the RCTs may have diminished the efficacy of screening with mammography and thereby reduced the extent of observed mortality reduction (*1,2*); (c) since the conduct of the RCTs, there have been numerous advances in mammographic equipment, technical imaging factors, quality assurance procedures, education of personnel, and mammographic interpretation performance (*1,3–6*), such that the mammography of the 1990s is demonstrably better than that done when the RCTs were conducted—advances that also may have caused the RCTs to underestimate the extent of currently achievable mortality reduction; and (d) the use of surrogate measures of screening efficacy, readily derived from modern service screening programs, demonstrates how the improved mammography of the 1990s can be expected to produce a greater degree of mortality reduction than that already demonstrated in the RCTs, thereby increasing the likelihood that modern mammography truly benefits screened women. Outcomes data from modern service screening programs also provide indicators of the frequencies with which abnormal screening interpretations, additional imaging evaluations, and screen-induced biopsies are performed in the real world, removed from the artificial conditions inherent in the design and conduct of the RCTs.

Many modern mammography service screening programs have published data in the peer review literature. These include: (a) the population-based screening program in Uppsala county, Sweden (*7*); (b) the province-wide Screening Mammography Program of British Columbia (SMPBC), Canada (*8*); (c) the University of California San Francisco (UCSF) screening program, which serves women throughout the San Francisco Bay Area (*9*); and (d) the X-Ray Associates of New Mexico screening program (*10*). These programs, which provide screening with mammography alone, were selected because they currently involve very large numbers of screening examinations (Uppsala, 127,515 examinations; SMPBC, 598,165 examinations; UCSF, 84,615 examinations; New Mexico, 104,371 examinations) and because they each collect extensive outcomes data including but not limited to linkage with comprehensive tumor registries in their respective geographic areas.

The outcomes data, displayed in tabular format throughout this article, are drawn from the UCSF program (*11,12*), utilizing updated results for all screening examinations performed through December 31, 1996. Because I have complete access to this source material, I can generate age-related data breakdowns that are not readily retrievable from any other service screening program or RCT.

Benefits of Screening With Modern Mammography

Since there has been controversy over the magnitude of the mortality reduction demonstrated by the RCTs among women

**Affiliation of author:* Department of Radiology, School of Medicine, University of California, San Francisco.

Correspondence to: Edward A. Sickles, M.D., Department of Radiology, UCSF/Mount Zion Medical Center, 2330 Post Street, #180, San Francisco, CA 94115.

ages 40–49 and since very few women aged 65+ have been studied in RCTs, surrogate measures of screening outcomes have been proposed, validated (using the same RCT data that indicate the presence of a mortality reduction), and widely used as alternative means to indicate the efficacy of screening (*7,12–22*). Two very powerful surrogate measures (i.e., those highly likely to predict reduced mortality) are tumor size and axillary lymph node status. Cancer stage, which is derived primarily from these two indicators, is the penultimate surrogate measure for screening efficacy; in fact, this measure is so valuable clinically that it is widely used in formulating treatment plans for breast cancer patients. Another useful measure is tumor grade. However, this cannot be evaluated readily in most service screening programs in the United States, because American pathologists too frequently omit grading data in their breast cancer reports (40% of cases in the UCSF program) and because many different pathologists perform grading assessments in the remaining cases, potentially introducing substantial subjective variation (*18,23*). A final surrogate measure of mortality reduction involves interval cancers—those cancers that are identified in the interval between screening examinations. Interval cancers grow more rapidly and have a poorer prognosis than screen-detected cancers (*7,15,24–27*); thus, a low interval cancer rate is a strong indicator of effective screening performance. However, the most valuable measure of interval cancers is the rate at which they occur in proportion to the rate at which cancers surface clinically in the absence of screening. Unfortunately, this measure is difficult to determine in the service screening setting, since there is no readily accessible control population of nonscreened women to provide the needed comparative data.

Surrogate measures of clinical efficacy are especially useful when employed in comparative studies—for example, in assessing the efficacy of different screening protocols (*17*). In this article, surrogate measures of mortality reduction will be used to compare the efficacy of screening women aged 40–49 and women aged 65+ with women of ages 50–64, the age range for which screening is widely accepted as being efficacious.

There is considerable evidence on the tumor size, lymph node status, and stage of cancers detected during modern service screening mammography. Indeed, these surrogate measures of mortality reduction appear to be as favorable for women ages 40–49 and 65+ as they are for women of ages 50–64 (see Table 1, which provides an update from the UCSF screening program, involving 72,145 examinations). Similar results also have been reported from the SMPBC, Uppsala, and New Mexico service screening programs (*7–10,20*). Thus, the surrogate-measure data strongly suggest that, since modern service screening is accepted as effectively reducing mortality among women in the 50–64 age group, it should also effectively reduce mortality among women of ages 40–49 and 65+.

There also is substantial evidence that the optimal screening interval for women aged 40–49 is one year, rather than the two-year interval used in most of the RCTs. Analysis of results from the Kopparberg portion of the Swedish two-county RCT, the Uppsala service screening program, and the Cincinnati Breast Cancer Detection Demonstration Project (another, albeit older service screening program) indicates that the lead time from screening women in their forties is substantially less than that from screening older women (*7,20,21,27*). Finally, as shown in Table 2, data from the UCSF service screening program demonstrate a substantial decline in sensitivity for screening women age 40–49 when the screening interval is increased from one year to two years, twice as large a decline as is observed for older women (*28*). This suggests that substantially more (poor-prognosis) interval cancers will occur if younger women are screened every two years rather than annually. Furthermore, the sensitivity for screening women age 40–49 at a one-year interval is equivalent to that of screening women age 50 and older at a two-year interval (*28*), an interval for the older cohort of women that already has been shown to produce statistically significant mortality reduction in the Swedish two-county RCT (*22*). Thus, these various lines of evidence all support the concept that if screening is recommended for women at ages 40–49, it should be done at an annual rather than a biennial interval.

It has been suggested that the slightly lower sensitivity for screening women aged 40–49 compared to older women may be a result of younger women's breasts being more radiographically dense. This argument is supported by the observation that the proportion of women with dense breasts is slightly higher at age 40–49 than it is in older women (*29,30*). However, the data on screening sensitivity from the UCSF program show that breast density did not influence the sensitivity of mammography in women aged 40–49; sensitivity was 90% for women with primarily dense breasts, compared with 88% for women with primarily fatty breasts (*28*). Similar findings also have been observed in the Swedish two-county RCT (*22*). A much more likely explanation for the slightly lower screening sensitivity in women age 40–49 is that rapid tumor growth rates among younger women result in more (poor-prognosis) interval cancers between screening examinations, as implied in the preceding discussion of optimal screening interval. This theory is further supported by the UCSF data, which show that screening sensitivity decreases with increasing tumor size, especially for women age 40–49 (*28*), suggesting that cancers in younger women grow more rapidly. Again, screening women age 40–49 at an annual rather than a biennial interval should result in sensitivity equivalent to that of screening older women at a two-year interval.

Table 1. Surrogate measures of breast cancer screening efficacy as a function of age at screening*

	Age 40–49	Age 50–64	Age 65+
Median size (invasive cancers)	12 mm	13 mm	12 mm
Nodal metastasis (invasive cancers)	15%	15%	14%
Stage 2 or higher cancers	19%	24%	18%

*Based on data from the UCSF service screening program, involving 425 cancers and 72,145 screening examinations.

Table 2. Sensitivity of initial screening mammography as a function of age at screening*

Follow-up interval	Age 40–49	Age 50+
1 year	87%	93%
2 years	73%	86%

*Based on data from the UCSF service screening program, derived from (*28*).

Although invasive breast cancer seems to grow more rapidly in women age 40–49 than in older women, does it really disseminate more frequently when very small in size (≤10 mm—i.e., detected primarily by screening)? Data from the Nijmegen population-based mammography screening program (1975–1990) for women with invasive cancers 10 mm or less show that 40% of women younger than age 50 had positive axillary lymph nodes, whereas only 20% of women age 50 and older had positive nodes (*31*). Dutch investigators have attributed these findings to presumed age-related differences in tumor biology, suggesting that, in younger women, cancers disseminate early in their evolution, whereas cancers in older women produce nodal metastasis more slowly. However, parallel evidence from the more modern UCSF service screening program is strikingly different. Among women with small (≤10 mm) invasive cancers, none aged 40–49 had positive axillary lymph nodes, and 6% of women aged 50–64 had positive nodes (*18*). Almost identical results have been reported for the New Mexico service screening program (*10*) and for the Swedish two-county RCT (*15*). Therefore, it appears likely that the Nijmegen observations do not serve as a basic indicator of tumor biology but simply are limited by relatively ineffective mammographic techniques and approaches (*18,32*). In conclusion, advancing the time of diagnosis for invasive cancers by screening does indeed appear to diminish the propensity for axillary lymph node metastasis and hence reduces the likelihood for mortality equally in women of ages 40–49 as in older women.

There are other benefits of screening women aged 40–49, apart from those indicated by the surrogate-marker evidence cited above. These range from the reassurance gained from knowledge that a screening examination was normal to the greater likelihood of being eligible for breast conservation therapy and of avoiding breast radiation therapy and chemotherapy when cancer is detected by screening, versus usual care. However, the outcomes data from modern service screening programs do not provide evidence to document such benefits, and so discussion of these benefits is beyond the scope of this article.

Risks of Screening With Modern Mammography

Several other measures of performance are also reported for service screening mammography, even though they do not appear to be reliable surrogates for breast cancer mortality. These include positive predictive value (PPV), biopsy yield of cancer, and specificity. These measures do, however, provide useful indicators of the frequency with which false-positive screening outcomes occur, thereby serving as surrogate measures for the risks (harms) of screening. It is important to note that PPV and biopsy yield are highly dependent on the prior probability of breast cancer, which increases steadily and substantially with advancing age, so that observed PPV and biopsy yield for women age 40–49 should be considerably lower than for older women.

Discussion of the nature and relative magnitudes of the risks of screening with mammography is also beyond the scope of this article. However, the outcomes data from modern service screening programs do provide relevant information, presented herein, on how some of these risks change with age. The risks of screening mammography should be considered separately for two specific populations of screened women.

The first population involves women recalled for additional noninterventional evaluation after abnormal mammography screening examinations. Outcomes data from modern service screening programs demonstrate that women age 40–49 are recalled for additional evaluation at approximately the same rate as women screened in later decades of life (*7,33*). When these data are examined by five-year age groupings, the same results are found (*34*). In the UCSF service screening program, there is essentially no difference in overall recall rate when comparing women age 40–49 with older women. Most recalled women will be found to have no clinically significant abnormalities. These women thus experience several negative outcomes of false-positive examination (anxiety, inconvenience, physical discomfort, cost). There also will be some women who eventually are found to have breast cancer, and because the incidence of breast cancer increases with advancing age, fewer women age 40–49 (than older women) will have true-positive examinations. Therefore, the PPV of screening will be lower for women age 40–49 than for older women (*7,19,35*). However, this age-dependent effect on true-positive examinations—that is, the prior probability of having breast cancer—is of very small overall magnitude, because less than 10% of recall examinations are true-positive examinations. Therefore, among women recalled for additional noninterventional evaluation after an abnormal screening examination, the overall risks are essentially age independent.

The second population involves women recalled for interventional evaluation (fine needle aspiration biopsy, core biopsy, or surgical biopsy) after abnormal diagnostic imaging examinations. Outcomes data from modern service screening programs demonstrate that women age 40–49 undergo these types of biopsy at approximately the same rate as women screened in later decades of life (*7,11,12,33*). When outcomes data are examined by five-year age groupings (and even by one year at a time), the same results are found (*36*). Most women undergoing biopsy will be found to have benign lesions. These women thus experience several negative outcomes of false-positive biopsy (anxiety, inconvenience, discomfort, scarring, cost), which are of greater magnitude than the risks described for recall examination. There also will be some women who eventually are found to have breast cancer, and because the incidence of breast cancer increases with advancing age, fewer women age 40–49 (than older women) will have true-positive biopsy. Therefore, the biopsy yield will be lower for women age 40–49 than for older women (*7,11,12,33,36*). However, this age-dependent effect on true-positive biopsy—again, the prior probability of having breast cancer—is of relatively small overall magnitude because only about one-third of biopsies are true-positive cases (*11,12*). Therefore, among women undergoing interventional evaluation after abnormal diagnostic imaging examinations, the overall risks are for the most part age independent.

Another point merits consideration concerning the biopsy of screen-detected lesions. In the United States, over the past five years, there has been a dramatic increase in the number of these lesions that undergo biopsy by percutaneous sampling (fine needle aspiration biopsy or core biopsy) rather than by surgical excision. Compared to surgical biopsy, percutaneous sampling is equally accurate but results in much less discomfort, produces

essentially no scarring, and is done at less than half the cost (*37–40*). When lesions undergoing percutaneous biopsy are found to be benign, in most cases surgical biopsy is averted, thereby resulting in substantially reduced morbidity. The cancer yield for surgical biopsy thus can be increased to between 60% and 75% (*7,15,39,41*). Because of the inherent advantages of percutaneous biopsy, the trend toward using this method rather than surgical biopsy for screen-detected lesions will probably continue at an accelerated rate as we proceed further into managed-care medicine.

One further useful piece of evidence can be derived from the UCSF service screening program. By comparing the outcomes from initial versus subsequent screening examinations, my colleagues and I demonstrated that the recall rate (frequency of abnormal screening interpretation) is substantially higher, the biopsy yield of cancer is considerably lower, and the surrogate measures of mortality reduction (tumor size, lymph node status, cancer stage) are less favorable for initial screening examinations than for subsequent examinations (*42*). Tables 3 and 4 present an update of UCSF screening program data on initial versus subsequent screening outcomes, demonstrating that the previously reported observations apply equally to women age 40–49 and to older women. It is important to note that ongoing service screening will involve many subsequent screening examinations but only one initial examination. Thus, outcomes data based either entirely or predominantly on initial screening will tend to underestimate the benefit and overestimate the risk of service screening.

Benefits and Risks of Screening With Modern Mammography, Applied to Populations of Women at Higher Than Average Risk for Breast Cancer

The RCTs were not designed to provide separate data on subpopulations of women at higher than average risk for developing breast cancer, and therefore no evidence on mortality reduction can be expected. However, outcomes data from the UCSF service screening program, using surrogate measures of screening performance, do provide the following indirect evidence on the benefits and risks of screening for women age 40–49 who have a strong or very strong family history of breast cancer: (a) the PPV of screening is higher in high-risk women age 40–49 than in the remainder of screened women in this age group (*35*), although it is likely that this simply is due to the increased incidence of breast cancer (greater prior probability of cancer) in these high-risk women rather than to an improved ability of screening to detect cancer in high-risk women; (b) the biopsy yield of cancer is higher in high-risk women age 40–49 (36%) than in the remainder of screened women in this age group (26%), again likely due to the increased incidence of breast cancer in high-risk women (greater prior probability of cancer); (c) the sensitivity of initial screening mammography is somewhat lower in high-risk women age 40–49 than in the remainder of initially screened women in this age group (*28*), likely due to a more rapid growth rate of cancers in younger high-risk women—these women do have a five-times greater risk of dying from breast cancer than younger average-risk women (*43*), suggesting that a greater proportion of cancers among younger high-risk women are aggressive and thus grow rapidly; (d) there are essentially no differences in the size, lymph node status, and stage of screen-detected breast cancers in comparing high-risk women age 40–49 with the remainder of screened women in this age group; and (e) had screening among women age 40–49 been limited to the 12% of these women at high risk by family history, this strategy would have detected only 19% of the extant cancers (*18,35*).

Table 3. Clinical outcomes of service screening mammography as a function of type of screening and age at screening*

	Initial		Subsequent	
	Age 40–49	Age 50+	Age 40–49	Age 50+
Screening examinations	1000	1000	1000	1000
Recalls for additional imaging	73	82	40	29
Biopsies performed	17	27	9	10
Cancers detected	4	12	3	4

*Based on data from the UCSF service screening program, involving 29,694 initial screening examinations and 42,451 subsequent screening examinations.

Table 4. Surrogate measures of breast cancer screening efficacy as a function of type of screening and age at screening*

	Initial		Subsequent	
	Age 40–49	Age 50+	Age 40–49	Age 50+
Median size (invasive cancers)	14 mm	15 mm	10 mm	11 mm
Nodal metastasis (invasive cancers)	17%	17%	13%	14%
Stage 2 or higher cancers	24%	26%	16%	19%

*Based on data from the UCSF service screening program, involving 425 cancers, 29,694 initial screening examinations, and 42,451 subsequent screening examinations.

Thus, the overall conclusion to be drawn from the UCSF experience is that, for the age range 40–49, modern service screening mammography appears to detect breast cancer somewhat less effectively in women at high risk for developing breast cancer, but that the accompanying increased incidence of breast cancer will increase the positive predictive value and biopsy yield, thereby improving the cost-effectiveness of screening these high-risk women. However, the more cost-effective strategy of screening only high-risk women will relinquish to usual-care detection more than 80% of the cancers in the entire age 40–49 population.

Directions for Future Research

There have been numerous advances in conventional mammography over the past 10 years, involving equipment, technical imaging factors, quality assurance procedures, education of personnel, and mammographic interpretation performance. Continued advances are expected as we enter the 21st century. There also is promising and very important research involving digital mammography, high-resolution breast ultrasound, magnetic resonance imaging, and isotope scanning of the breast. Among these imaging techniques, digital mammography may provide increased sensitivity and/or specificity when used for breast cancer screening. All techniques may permit increased sensitivity

and/or specificity in the "diagnostic" setting (i.e., providing noninterventional evaluation of screen-detected abnormalities).

In contrast to breast imaging, which has undergone (and continues to undergo) many improvements, very little change has occurred in the practice of breast physical examination, other than the realization that it appears to be more accurate when performed with diligence by specially trained practitioners (*44*). Unfortunately, there currently is little enthusiasm either within or outside the medical community for improving the current state of breast physical examination in the United States. Two approaches that are likely to reap considerable benefit are (a) the recruitment, training, and deployment of large numbers of paramedical personnel to perform breast physical examination in screening centers and (b) federal legislation mandating quality assurance practices for breast physical examination, to parallel the provisions of the Mammography Quality Standards Act of 1992 (which has resulted in considerably improved delivery of high-quality mammography services).

The National Cancer Institute has funded a multisite Breast Cancer Surveillance Consortium, which is currently collecting outcomes data from more than one million women on many aspects of modern breast cancer screening. This research will provide valuable direction into methods of improving breast cancer screening in the United States. However, there is urgent need to go beyond this effort by creating a national cancer registry to permit collection of meaningful outcomes data for all American women. To be truly effective, such a cancer registry must permit low-cost data linkage by individual breast cancer screening practices, so that complete rather than partial outcomes data are available to service providers for the purpose of continuous quality improvement.

References

(*1*) Sickles EA, Kopans DB. Deficiencies in the analysis of breast cancer screening data [editorial]. J Natl Cancer Inst 1993;85:1621–4.

(*2*) Feig SA. Determination of mammographic screening intervals with surrogate measures for women aged 40–49 years [editorial]. Radiology 1994; 193:311–4.

(*3*) Conway BJ, McCrohan JL, Reuter FG, Suleiman OH. Mammography in the eighties. Radiology 1990;177:335–9.

(*4*) Conway BJ, Suleiman OH, Reuter FG, Antonsen RG, Slayton RJ. National survey of mammographic facilities in 1985, 1988, and 1992. Radiology 1994;191:323–30.

(*5*) McLelland R, Hendrick RE, Zinninger MD, Wilcox PA. The American College of Radiology Mammography Accreditation Program. AJR Am J Roentgenol 1991;157:473–9.

(*6*) Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases [published erratum appears in Radiology 1992;184:878]. Radiology 1992;184:39–43.

(*7*) Thurfjell EL, Lindgren JA. Population-based mammography screening in Swedish clinical practice: prevalence and incidence screening in Uppsala county. Radiology 1994;193:351–7.

(*8*) Burhenne LW, Hislop TG, Burhenne HJ. The British Columbia Mammography Screening Program: evaluation of the first 15 months. AJR Am J Roentgenol 1992;158:45–9.

(*9*) Sickles EA, Weber WN, Galvin HB, Ominsky SH, Sollitto RA. Mammography screening: how to operate successfully at low cost. Radiology 1986; 160:95–7.

(*10*) Linver MN, Paster SB. Mammography outcomes in a practice setting by age: prognostic factors, sensitivity, and positive biopsy rate. Monogr J Natl Cancer Inst, 1997;22:113–117.

(*11*) Sickles EA, Ominsky SH, Sollitto RA, Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations. Radiology 1990;175:323–7.

(*12*) Sickles EA. Auditing your practice. In: Kopans DB, Mendelson EB, editors. Syllabus: a categorical course in breast imaging. Oak Brook (IL): Radiological Society of North America, 1995:81–91.

(*13*) Tabar L, Larsson LG, Andersson I, et al. Breast-cancer screening with mammography in women aged 40–49 years. Int J Cancer 1996;68:693–9.

(*14*) Day NE. Quantitative approaches to the evaluation of screening programs. World J Surg 1989;13:3–8.

(*15*) Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.

(*16*) Bassett LW, Hendrick RE, Bassford TL, et al. Quality determinants of mammography: clinical practice guideline no. 13. AHCPR Publ No. 95-0632. Rockville (MD): DHHS, PHS, AHCPR, 1994.

(*17*) Feig SA. Determination of mammographic screening intervals with surrogate measures for women aged 40–49 years [editorial]. Radiology 1994; 193:311–4.

(*18*) Curpen BN, Sickles EA, Sollitto RA, Ominsky SH, Galvin HB, Frankel SD. The comparative value of mammographic screening for women 40–49 years old versus women 50–64 years old. AJR Am J Roentgenol 1995; 164:1099–103.

(*19*) Burhenne LW, Burhenne HJ, Kan L. Quality-oriented mass mammography screening. Radiology 1995;194:185–8.

(*20*) Thurfjell EL, Lindgren JA. Breast cancer survival rates with mammographic screening: similar favorable survival rates for women younger and those older than 50 years. Radiology 1996;201:421–6.

(*21*) Tabar L, Fagerberg G, Day NE, Holmberg L. What is the optimum interval between screening examinations? An analysis based on the latest results of the Swedish two-county breast cancer screening trial. Br J Cancer 1987; 55:547–51.

(*22*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age: new results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.

(*23*) Duffy SW, Tabar L, Fagerberg G, Gad A, Grontoft O, South MC, et al. Breast screening, prognostic factors and survival—results from the Swedish two county study. Br J Cancer 1991;64:1133–8.

(*24*) DeGroote R, Rush BF Jr, Milazzo J, Warden MJ, Rocko JM. Interval breast cancer: a more aggressive subset of breast neoplasias. Surgery 1983;94: 543–7.

(*25*) Andersson I. What can we learn from interval carcinomas? Recent Results Cancer Res 1984;90:161–3.

(*26*) Frisell J, Eklund G, Hellstrom L, Somell A. Analysis of interval breast carcinomas in a randomized screening trial in Stockholm. Breast Cancer Res Treat 1987;9:219–25.

(*27*) Moskowitz M. Breast cancer: age-specific growth rates and screening strategies. Radiology 1986;161:37–41.

(*28*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276:33–8.

(*29*) Kopans DB. Conventional wisdom: observation, experience, anecdote, and science in breast imaging. AJR Am J Roentgenol 1994;162:299–303.

(*30*) Stomper PC, D'Souza DJ, DiNitto PA, Arredondo MA. Analysis of parenchymal density on mammograms in 1353 women 25–79 years old. AJR Am J Roentgenol 1996;167:1261–5.

(*31*) Peer PG, Holland R, Hendriks JH, Mravunac M, Verbeek AL. Age-specific effectiveness of the Nijmegen population-based breast cancer screening program: assessment of early indicators of screening effectiveness. J Natl Cancer Inst 1994;86:436–41.

(*32*) Kopans DB. Mammographic screening for breast cancer [editorial]. Cancer 1993;72:1809–12.

(*33*) Olivotto IA, Warren L. 1995–96 annual report. Vancouver (BC): Screening Mammography Program of British Columbia, 1996.

(*34*) Burhenne HJ, Burhenne LW, Goldberg F, Hislop TG, Worth AJ, Rebbeck PM, et al. Interval breast cancers in the Screening Mammography Program of British Columbia: analysis and classification. AJR Am J Roentgenol 1994;162:1067–71.

(*35*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.

(*36*) Kopans DB, Moore RH, McCarthy KA, Hall DA, Hulka CA, Whitman GJ, et al. The positive predictive value of breast biopsy performed as a result of mammography: there is no abrupt change at age 50 years. Radiology 1996; 200:357–60.

(*37*) Parker SH, Burbank F, Jackman RJ, Aucreman CJ, Cardenosa G, Cink TM, et al. Percutaneous large-core breast biopsy: a multi-institutional study. Radiology 1994;193:359–64.

(*38*) Brenner RJ, Fajardo L, Fisher PR, Dershaw DD, Evans WP, Bassett L, et al. Percutaneous core biopsy of the breast: effect of operator experience and

number of samples on diagnostic accuracy. AJR Am J Roentgenol 1996; 166:341–6.
(*39*) Azavedo E, Svane G, Auer G. Stereotactic fine-needle biopsy in 2594 mammographically detected non-palpable lesions. Lancet 1989;1:1033–6.
(*40*) Liberman L, Fahs MC, Dershaw DD, Bonaccio E, Abramson AF, Cohen MA, et al. Impact of stereotaxic core breast biopsy on cost of diagnosis. Radiology 1995:195;633–7.
(*41*) Logan-Young WW, Janus JA, Destounis SV, Hoffman NY. Appropriate role of core breast biopsy in the management of probably benign lesions. Radiology 1994:190:313.
(*42*) Frankel SD, Sickles EA, Curpen BN, Sollitto RA, Ominsky SH, Galvin HB. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. AJR Am J Roentgenol 1995; 164:1107–9.
(*43*) Calle EE, Martin LM, Thun MJ, Miracle HL, Heath CW Jr. Family history, age, and risk of fatal breast cancer. Am J Epidemiol 1993;138:675–81.
(*44*) Baines CJ, Miller AB, Bassett AA. Physical examination. Its role as a single screening modality in the Canadian National Breast Screening Study. Cancer 1989;63:1816–22.

Outcomes of Modern Screening Mammography

*Karla Kerlikowske, John Barclay**

The University of California, San Francisco, Mobile Mammography Screening Program is a low-cost, community-based breast cancer screening program that offers mammography to women of diverse ethnic backgrounds (36% nonwhite) in six counties in northern California. Analysis of data collected on approximately 34,000 screening examinations from this program shows that the positive predictive value and sensitivity of modern screening mammography to be lower for women aged 40 to 49 years compared to women aged 50 and older. This lower performance is due to the lower prevalence of invasive breast cancer in younger women and possibly to age differences in breast tumor biology. Because of this lower performance, women in their forties may be subjected to more of the negative consequences of screening, which include additional diagnostic evaluations and the associated morbidity and anxiety, the potential for detecting and surgically treating clinically insignificant breast lesions, and the false reassurance resulting from normal mammographic results. Since the evidence is not compelling that the benefits of mammography screening outweigh the known risks for women aged 40 to 49 years, women considering mammography screening should be informed of the risks, potential benefits, and limitations of screening mammography, so that they can make individualized decisions based on their personal risk status and utility for the associated risks and potential benefits of screening. [Monogr Natl Cancer Inst 1997:22:105–111]

Randomized controlled screening mammography trials have not conclusively demonstrated a reduction in breast cancer mortality for women aged 40 to 49 years, at least not for the first seven to nine years after the initiation of screening (*1–3*). If screening mammography is effective in reducing breast cancer deaths among women aged 40 to 49 years, the reduction in deaths does not occur for at least a decade following the initiation of screening, and it appears to be smaller than the reduction observed in women aged 50 and older, resulting in a small absolute benefit from screening younger women (*4*). Screening mammography may be less effective for women aged 40 to 49 years in part because mammography is less sensitive in younger women. Some have argued that with the improvement in the quality of modern mammography, specifically its sensitivity, the results reported from previous randomized controlled trials are not generalizable to women today. However, the question remains whether the performance of modern screening mammography has improved for younger women. We review evidence of the performance of modern screening mammography from the University of California, San Francisco (UCSF), Mobile Mammography Screening Program and discuss possible explanations as to why the performance may differ in younger compared to older women. We also present the potential negative consequences of performing widespread screening mammography among young women based on the performance of modern screening mammography. Lastly, we discuss the potential association between widespread screening mammography and the decrease in breast cancer mortality in the United States reported in 1992 and 1993 (*5*).

Definitions

There are several parameters used to evaluate the performance of screening mammography. The most widely used parameters are the percent of all screening examinations that have abnormal results (or simply, "percent abnormal"), the positive predictive value (PPV) of mammography, the yield of breast biopsy, and the sensitivity of mammography. For our purposes here, an "abnormal" screen is a screening examinations that requires any additional tests beyond the standard two-view examination, be it additional mammographic views, ultrasound, clinical breast exam, fine needle aspiration, or breast biopsy. The PPV of screening mammography is the percent of women with abnormal screening examinations who are subsequently diagnosed with breast cancer. The yield of breast biopsy is the percent of women who undergo breast biopsies that result in a diagnosis of breast cancer. The sensitivity of mammography is calculated as the number of true positive examinations divided by the number of true positive plus false negative examinations. A true positive examination is defined as an abnormal mammographic examination (of a specified breast) that is performed within 13 months prior to the date of a biopsy with a diagnosis of breast cancer and a false negative examination is defined as a normal mammographic examination (of a specified breast) that was performed within 13 months prior to the date of a biopsy with a diagnosis of breast cancer that was presented clinically.

Breast cancer is defined as any invasive cancer or ductal carcinoma *in situ* (DCIS). DCIS is a proliferation of cells with malignant features that is confined within the mammary ducts. DCIS is a "nonobligate" premalignant lesion—that is, it has the potential to progress to invasive cancer but does not always automatically do so. DCIS lesions are easier to detect because

**Affiliations of authors:* K. Kerlikowske, Department of Epidemiology and Biostatistics, University of California, San Francisco, and General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco; J. Barclay, Department of Epidemiology and Biostatistics, University of California, San Francisco.

Correspondence to: Dr. Karla Kerlikowske, San Francisco Veterans Affairs Medical Center, General Internal Medicine Section, 111A1, 4150 Clement Street, San Francisco, CA 94121.

See "Note" following "References."

they usually present as microcalcifications (*6,7*) on mammography, whereas invasive cancer usually presents as noncalcified masses. Of those DCIS lesions that progress to invasive cancer, most do so slowly, taking five to 10 years to develop into invasive cancer (*8–12*). Since the identification and growth rates of DCIS are different than for invasive breast cancer and because the proportion of mammographically detected cancer that is DCIS varies with age (*13*), data on the parameters defined above are presented separately for invasive cancer and all breast outcomes (invasive cancer and DCIS).

Performance of Modern Screening Mammography

The percent abnormal of first screening examinations increases with age from 6.4% in women aged 40 to 49 years to 8.0% in women aged 60 to 69 years (Table 1). The PPV of mammography also increases with age, with women aged 50 to 59 years having about a twofold higher PPV of mammography than women aged 40 to 49 years (Table 1). This means for every 100 women in their forties with abnormal mammographic results, about 2.5 will have invasive cancer, compared with 6.3 and 12.2 per 100 women in their fifties and sixties, respectively. The PPV of mammography is somewhat higher for all ages of women when all breast cancer outcomes are considered but still remains low for women aged 40 to 49 years, with only 4.6 cancers for every 100 abnormal first screening examinations. The PPV of mammography we report for first screening mammography is consistent with that reported by the Canadian National Breast Cancer Screening Study for women aged 40 to 49 years (4.4%) (*14*) and somewhat higher than that reported for modern screening mammography by a recent British Columbia study (2.0%) (*15,16*).

Table 1. Performance of first and subsequent screening mammography

	Age (years)		
	40 to 49	50 to 59	60 to 69
First screening			
Abnormal exams (%)	6.4	6.8	8.0
Breast cancers/1,000 exams	3	6	12
(95% CI)	(2, 4)	(5, 8)	(9, 16)
PPV mammography			
Average-risk*			
Invasive cancer only (%)	2.6	6.3	12.2
(95% CI)	(1.7, 4.0)	(4.4, 9.0)	(9.1, 16.1)
All breast cancer (%)	4.6	9.0	14.9
(95% CI)	(3.3, 6.3)	(6.6, 12.0)	(11.4, 19.1)
Family history of breast cancer†			
All breast cancer (%)	9.2	16.4	12.1
(95% CI)	(4.3, 17.8)	(9.1, 27.3)	(4.0, 29.1)
Subsequent screening*‡			
Abnormal exams (%)	3.2	2.5	2.0
Breast cancers/1,000 exams	2	4	3
(95% CI)	(1, 3)	(2, 6)	(1, 5)
PPV mammography			
Average-risk*			
All breast cancer (%)	6.0	16.0	12.5
(95% CI)	(3.0, 11.5)	8.9, 26.7)	(4.7, 27.6)

*Data from UCSF Mobile Mammography Screening Program, 1985–1996. Excludes women with a history of breast cancer or mastectomy, palpable mass by history or physical exam, or family history of breast cancer. All breast cancer includes invasive cancer and ductal carcinoma *in situ*.

†Defined as at least one first-degree relative (mother, sister, or daughter) with breast cancer.

‡Includes only second screening examinations.

The observed increase in PPV with increasing age is most likely due to the higher prevalence of breast cancer in older women. The incidence of breast cancer increases approximately 1.5-fold every 10 years from age 40 to age 70, with approximately 76% of all invasive breast cancers diagnosed after age 50 (*17*). Thus, even though women aged 50 and older only comprise 30% of all women in the United States (*18*), the majority of breast cancer is detected at or after age 50. Our results reflect this increasing incidence, as the number of breast cancers detected per 1,000 first screening examinations doubles with each 10-year increase in age (Table 1).

In addition to age, a family history of breast cancer affects the PPV of mammography. The relative risk of breast cancer is two to three times higher in women who have had a first-degree relative diagnosed with breast cancer (*19,20*). The higher risk of breast cancer among women with a family history of breast cancer increases the prevalence of breast cancer in these women, and consequently the PPV (Table 1). This is particularly true for women aged 40 to 49 years and women aged 50 to 59 years with a family history, since the relative increase in risk of breast cancer, compared to women without a family history, is higher for women under 60 than for those aged 60 and older (*20*).

The percent abnormal and the PPV of mammography is also affected by the percentage of the population being screened for the first time. The percent abnormal for subsequent screening examinations is lower for all ages of women, but it decreases with increasing age (Table 1). The lower percent abnormal on subsequent screening is primarily due to fewer examinations being interpreted as abnormal when first-screening films are available for comparison. This results in higher PPVs for mammography on subsequent screening examinations for women of all ages (Table 1). Of note, however, is that the PPV for subsequent screening mammography for women aged 40 to 49 years is still low (6%) and less than both the PPV of subsequent screening mammography for women age 50 to 59 years (16%) and the PPV of first screening mammography for women ages 50 to 59 (9%).

Another measure of the performance of modern screening mammography is the yield of cancer diagnosed per breast biopsy performed. The number of biopsies per 1,000 exams increases with age, as does the yield of cancer (Table 2). Therefore, even though more biopsies are performed in older women, more cancer is detected per biopsy performed. For women aged 40 to 49 years, one in five biopsies will have invasive cancer or DCIS and only one in 10 will have invasive cancer (Table 2). The yield of cancer is greater in older women, for whom about one in three biopsies will have invasive cancer or DCIS, and about one in four will have invasive cancer. The lower yield of invasive cancer in younger women is due to the lower incidence of breast cancer in these women and the higher proportion of mammographically detected cancer being DCIS (Table 2).

Many may feel that the low PPV of modern mammography, which results in many abnormal examinations that do not result in a diagnosis of breast cancer (false-positive), is acceptable as long as cancer does not go undetected. Therefore, the critical question is, How sensitive is mammography in detecting breast cancer among women who have the disease? Studies of modern

Table 2. Results of breast biopsies in women after first screening mammography*

	Age (years)		
	40 to 49	50 to 59	60 to 69
Breast biopsies/1,000 exams	15	19	28
(95% CI)	(13, 17)	(17, 23)	(24, 34)
Breast biopsy interpretation			
Invasive cancer (%)	56	71	82
DCIS† (%)	44	30	18
Breast cancer/biopsy			
Invasive cancer only (%)	11	22	34
(95% CI)	(7, 16)	(16, 30)	(26, 43)
All breast cancer‡ (%)	20	32	42
(95% CI)	(14, 26)	(24, 40)	(34, 51)

*Data from UCSF Mobile Mammography Screening Program, 1985–1996. Excludes women with a history of breast cancer or mastectomy, palpable mass by history or physical exam, or family history of breast cancer.

†Ductal carcinoma *in situ*.

‡All breast cancer includes invasive cancer and ductal carcinoma *in situ*.

screening mammography (*16,21–26*) report overall sensitivities of screening mammography (71.1% to 91.5%) similar to those published for clinical trials (*27,28*). Two studies report the sensitivity of mammography by age, and they show that sensitivity is still lower for women less than age 50 years (63% and 80%) compared to women aged 50 and older (89% and 94%) (*16,23*). A recent study (*21*) that evaluated the sensitivity of modern screening mammography by decade of age showed that the sensitivity of mammography to detect invasive breast cancer is still lower among women aged 40 to 49 years compared with women aged 50 and older (75% versus 92%). An updated analysis of these data (*21*) found similar results (Table 3). Conventional thinking has been that the lower sensitivity is due to the lower fat content of younger women's breasts, making them less radiolucent on film screen mammography (and thus obscuring small tumors) than those of older women. However, two recent studies have shown that the sensitivity of mammography does not vary according to breast density among younger women (*21,28*). Rather, the lower sensitivity in younger women is more likely a result of rapid tumor growth rates (*21*).

Even though the absolute benefit of screening women aged 40 to 49 years is small (*4,18,29*) and the ability to detect invasive cancer is less in comparison to older women, why not do it anyway? The main reasons to not recommend mass screening when the benefits of the screening test are uncertain, or of small benefit, are the following: 1) the burden of unnecessary workups of false-positive examinations with associated morbidity, anxiety, and cost; 2) the potential to detect lesions that may be clinically insignificant yet are treated anyway; and 3) the false reassurance resulting from a normal screening test result.

Table 3. Sensitivity of first screening mammography*

	Age (years)		
Sensitivity	40 to 49	50 to 59	60 to 69
Invasive cancer only (%)	78.0	90.9	89.8
(95% CI)	(62.0, 88.9)	(77.4, 97.0)	(77.0, 96.2)
DCIS (%)	100	100	100
All breast cancer† (%)	84.7	93.0	91.2
(95% CI)	(72.5, 92.4)	(82.2, 97.7)	(80.0, 96.7)

*Data from UCSF Mobile Mammography Screening Program, 1985–1994.

†All breast cancer includes invasive cancer and ductal carcinoma *in situ*.

False-Positive Examinations

Nationwide, about 11% of all screening examinations are read as abnormal (range 3–57%), with the average PPV of mammography for women aged 40 to 49 years being about twice as low as that for women aged 50 and older (2.0 versus 4.7) (*30*). Even at institutions with well-trained, full-time mammographers, about 6% of first screening mammography examinations are read as abnormal and the PPV of mammography is low (*13*). One consequence of the low PPV of mammography is an increase in the number of diagnostic evaluations. Since the PPV of mammography is low in women aged 40 to 49 years, these women may be subjected to the greatest harm, since they will undergo the greatest number of diagnostic tests to find the fewest cancers. For example, among 100 average-risk women aged 40 to 49 years with an abnormal first screening examination, about 95 do not have cancer (Table 1) and must undergo further diagnostic evaluation, which may include tests such as clinical breast examination, additional mammography, ultrasound, needle aspiration, or excisional biopsy. On average, approximately 1.5 to two additional diagnostic tests are performed per abnormal screening examination (*13,31*). Because many mammographic abnormalities are nonpalpable, needle localization biopsy is often required. Although risk is low, there are complications associated with biopsies, such as hematomas, infection, and scarring, and from wire localization itself, complications such as vasovagal reactions (7%) and, in rare cases, prolonged bleeding (1%) and extreme pain (1%) (*32*). In addition, a substantial proportion of women have increased anxiety about breast cancer, compared to women with normal mammographic results, even after learning they do not have cancer (*33–36*). Twenty-nine percent have persistent anxiety 18 months after an abnormal mammographic result compared to women with a normal mammographic result (13%), and women who undergo breast biopsies have especially high anxiety (*33*). However, such anxiety does not appear to interfere with subsequent adherence to screening. In contrast, women who do not have anxiety about breast cancer, or women who have decreased anxiety about breast cancer after undergoing screening mammography, are less likely to obtain subsequent annual mammography. Lastly, some women may be wrongly labeled as being at higher risk of breast cancer as a result of having a false-positive mammographic examination which may affect recommendations for subsequent screening and insurance status.

Assuming a high level of mammography performance (*13*), if 10,000 average-risk women aged 40 to 49 years undergo screening mammography for the first time, approximately 640 will have an abnormal finding requiring some additional test (including 150 biopsies); 30 will have cancer, 17 of which will be invasive cancer and 13 DCIS. In comparison, if 10,000 average-risk women aged 50 to 59 years undergo screening mammography for the first time, approximately 680 women will have an abnormal finding requiring some additional test (including 188 biopsies); 60 will have cancer, 42 of which will be invasive and 18 DCIS (Table 4). Thus, women aged 40 to 49 years will undergo a similar number of biopsies to diagnose half as many

Table 4. Comparison of first screening results by age*

	Age (years)		
	40 to 49	50 to 59	60 to 69
Number screened	10,000	10,000	10,000
Number of abnormals	640	680	800
Number of tests†	1280	1360	1600
Number of biopsies	150	188	286
Number of invasive cancers	17	42	98
Number of DCIS‡	13	18	22

*Data from UCSF Mobile Mammography Screening Program, 1985–1996. Excludes women with a history of breast cancer or mastectomy, palpable mass by history or physical exam, or family history of breast cancer.

†Includes clinical breast examination, additional mammography, ultrasound, fine needle aspiration, and excisional biopsy.

‡Ductal carcinoma *in situ*.

breast cancers compared with women aged 50 to 59 years, will have two times as many diagnostic tests (43 versus 23) for every DCIS or invasive cancer diagnosed, and will have 2.5 times as many tests (75 versus 32) for every invasive cancer diagnosed. The lower yield of cancer per breast biopsy and higher number of diagnostic tests per cancer detected in women aged 40 to 49 years is a result of the lower incidence of breast cancer in these women.

When speaking with women who are considering screening mammography, health practitioners should inform them of both the potential benefits and harms of screening. For a 40 year old woman who elects to be screened annually for ten years (i.e., has ten mammographic examinations in ten years), she should be informed she has a 30% chance of having at least one abnormal screening examination that will require a diagnostic work-up, a 28% chance of at least one false-positive examination, and a 7.5% chance of undergoing at least one breast biopsy (Table 5). For a 50-year-old woman who elects to be screened annually for ten years, she should be informed she has a 26% chance of having at least one abnormal screening examination that will require a diagnostic work-up, a 23% chance of at least one false-positive examination, and a 10.4% chance of undergoing at least one breast biopsy. For all women, irrespective of age, the chance of an abnormal test and false-positive test is greater than the risk of breast cancer (Table 5). However, for younger women, the risk of a false-positive test is the highest because the incidence of breast cancer is lower in these women. It is important to emphasize that these numbers are based on abnormal rates for first screening and subsequent screening (Table 1) that assume high-quality screening mammography. Thus, the numbers presented may be a conservative estimate of the risk of an abnormal examination, a false-positive result, and a breast biopsy over ten years of screening. Results from the Canadian National Breast Screening Study of women aged 40 to 49 years are comparable to the results presented in Table 5, with a 16.9% five-year cumulative risk of being recalled for evaluation of an abnormal mammographic examination after five screening exams and 11.8% after three exams (Personal communication from Anthony Miller, Ph.D.). In contrast, a study of women aged 40 to 69 years in a health maintenance organization has reported a 21% 10-year cumulative risk of a false-positive exam after only three screening examinations (*37*).

Table 5. Risk of at least one abnormal mammographic exam, false-positive exam, and breast biopsy if screened annually for 10 years*

	Age (years)		
Risk	40	50	60
Abnormal exam*	30%	26%	23%
False-positive exam*	28%	23%	20%
Biopsy*	7.5%	10.4%	10.4%
Breast cancer†	1.5%	2.4%	3.4%

*Calculations based on results presented in Table 1 and 2.

†Risk of breast cancer in the next 10 years for a 40-, 50-, and 60-year-old woman (*17*).

Overdiagnosis of Clinically Insignificant Lesions

The point of screening is to discover potentially fatal cancers early enough to prevent death. However, screening tends to discover cancers that may never have produced symptoms. The best example of this is DCIS. The natural history of DCIS is unknown, in particular, the natural histories of many small mammographically detected DCIS lesions. Numerous studies have shown that only 15% to 25% of DCIS lesions progress to invasive cancer over 5 to 10 years (*38–42*) and maybe as few as 7% (*12*). Of breast cancers detected by screening mammography in average-risk women aged 40 to 49 years, approximately 44% are DCIS, compared to 20%–30% of those detected in women aged 50 and older (Table 2). Given that the natural history of DCIS is unknown, the current clinical dilemma lies in not being able to distinguish which lesions will progress to invasive cancer. Thus, screening mammography may be benefiting some women through early detection of potentially fatal breast cancers, while it is potentially harming other women through detection of clinically insignificant lesions that, for lack of good prognostic indicators, are almost always treated surgically (*43*).

Potential for False Reassurance

Of 100 women aged 40 to 49 years with invasive breast cancer, about 22 will go undetected by screening mammography, compared with 9 of 100 women aged 50 to 59 years with invasive cancer (Table 3). This means potentially 22 women aged 40 to 49 years with invasive breast cancer will be told their screening examination is normal and may be falsely reassured that they do not have breast cancer and thus not seek medical attention for breast symptoms. For women who do not have breast cancer, they may also be reassured by having a normal screening examination that they do not have breast cancer. The annual risk of breast cancer for a 40-year-old woman is about 1 in 625 (*17*); having a normal screening examination decreases her risk to about 1 in 2500 (*44*). Although the very low risk of breast cancer after a normal screening examination may reassure women that they do not have breast cancer, the risk of breast cancer *before* mammography is already quite low. The need for reassurance from mammography may not be necessary if women in their forties understood that the risk of breast cancer prior to mammography is already very low (*45*). Thus, screening mammography is not justified solely to reassure women that breast cancer is not present; moreover, women should be informed that cancer may go undetected by mammography.

Decreased Breast Cancer Mortality in the United States—Is It From Screening?

Recently published data from the National Cancer Institute (NCI) show that among white women from 1989 to 1993, breast cancer mortality has decreased 8% in women 40–49, 9% in women 50–59, and 5% in women 60–69 (*5*). Proponents of screening mammography for women aged 40 to 49 years have suggested that this decrease is due to the improvement in, and widespread use of, modern screening mammography (*5*). There are many reasons why breast cancer mortality may be decreasing in the United States, however, including more widespread use of adjuvant therapy, improved detection by mammography, a shift in the risk factors for breast cancer in the population, earlier reporting of breast symptoms, and cohort effect. No randomized controlled trial has been conducted to test whether modern mammography results in a reduction in breast cancer mortality among women aged 40 to 49 years.

An indirect way to examine whether the increase in modern mammography utilization has affected breast cancer mortality is to look at NCI's population-based Surveillance, Epidemiology, and End Results (SEER) tumor registry data (*17*) to see if there has been a decrease in the incidence of late-stage disease. Specifically, if mammography accounts for the observed decrease in breast cancer mortality, then screening should advance the time of diagnosis and result in a lower rate of breast cancer cases having lymph node involvement. In other words, a lower rate of lymph node involvement would result in a decrease in breast cancer mortality, since lymph node involvement has the greatest impact on breast cancer survival.

In examining the population-based SEER tumor registry data for white women (*17*), we considered all DCIS lesions and invasive tumors that were less than 20 mm without associated positive lymph nodes to be consistent with screening or early-stage cancer; invasive tumors 20 mm or larger or those tumors associated with positive lymph nodes regardless of tumor size were considered to be inconsistent with screening or late-stage cancer (*46*). Among women in their fifties and sixties, with the increase in the rate of early-stage disease, there has been a persistent decrease in late-stage disease since 1986 (Figure 1a and 1b). Therefore, it appears as if there has been a shift from more advanced-stage disease to earlier-stage disease, such that the rate of tumors consistent with screening is higher than the rate of tumors not consistent with screening. The increase in early-stage disease has been tied to the dramatic increase in use of screening mammography (*47,48*). Therefore, the six-year decline in late-stage disease for women aged 50 to 59 and 60 to 69 years suggests that the decline in breast cancer mortality observed in 1992 and 1993 may be due, in part, to screening mammography. Other likely explanations for the decline in late-stage disease could be earlier reporting of breast symptoms or cohort effect.

For women aged 40 to 49 years, the rate of tumors not consistent with screening was similar in 1983 as in 1991 (Fig. 1c). Not until 1992 was there a decline in tumors not consistent with screening or late-stage disease for women aged 40 to 49 years. Therefore, although modern mammography has resulted in an increase in breast cancer cases consistent with screening among younger women, it has not resulted in a shift from more advanced-stage disease to early-stage disease. Thus, given that there has not been a persistent decline in late-stage disease, it is less likely that the decrease in breast cancer mortality observed in 1992 and 1993 among white women aged 40 to 49 years is due to screening mammography. As noted above, there are many reasons why breast cancer mortality may have declined in the United States, including improved breast cancer treatment. The United Kingdom has also reported a 9.8% decline in breast cancer mortality among women aged 40 to 49 years between 1989 and 1994, despite the fact that younger women do not undergo regular screening mammography, since mass screening is not recommended for women under age 50 (*49*). Taken together, these results suggest that the decline in breast cancer mortality among women aged 40 to 49 years is less

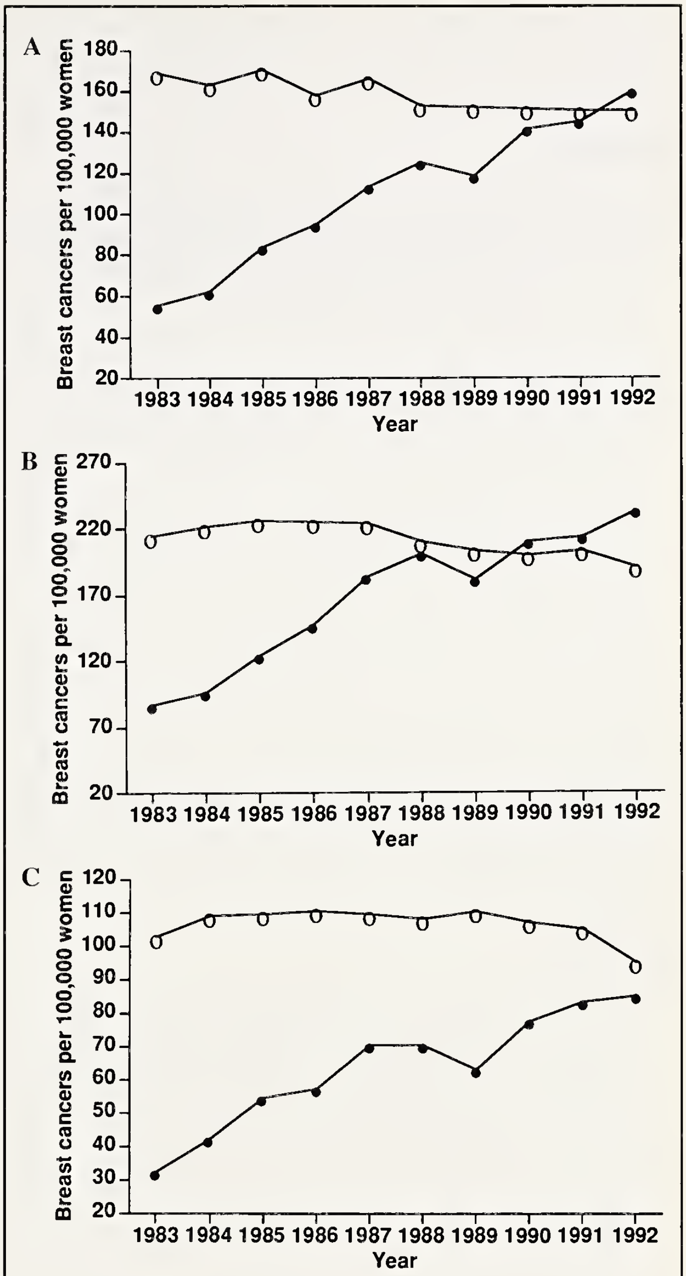

Fig. 1. Population-based SEER data showing incidence of early- versus late-stage disease among white women by decade of age. A) Women aged 50 to 59 years, B) Women aged 60 to 69 years, C) Women aged 40 to 49 years. ● = Early; ○ = Late.

likely due to early detection from screening mammography and more likely due to other reasons, such as improved breast cancer treatment.

Conclusion

There are associated risks with undergoing screening mammography, including additional diagnostic evaluations and the associated morbidity and anxiety, the potential for detecting and surgically treating clinically insignificant breast lesions, and the false reassurance resulting from a normal mammographic result. Before mass screening is recommended to healthy persons, the benefits of the intervention should be proven to clearly outweigh the risks. Given that the small absolute benefit (*4*) does not clearly outweigh the known risks, health practitioners should instead inform women of the risks, potential benefits, and limitations of screening mammography, so that each woman can make an individualized decision based on her personal risk status and utility for the associated risks and potential benefits of screening. Women who request or are offered screening mammography should be informed of the following: 1) their age-specific risk of breast cancer, 2) the chance of undergoing a diagnostic procedure, 3) the chance of a false negative, and 4) the evidence that screening mammography reduces the risk of death among screened women in their age group. In addition, health practitioners need to assist women in understanding what factors might influence their choice to undergo or not undergo screening, such as their attitude toward pain, risk, and inconvenience (*50*).

References

(*1*) Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.
(*2*) Elwood JM, Cox B, Richardson AK. The effectiveness of breast cancer screening by mammography in younger women [published errata appear in Online J Curr Clin Trials 1993; Doc No. 34 and 1994; Doc No. 121]. Online J Curr Clin Trials 1993; Doc No. 32.
(*3*) Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. Med J Aust 1995;162:625–9.
(*4*) Kerlikowske K. Efficacy of screening mammography among women aged 40 to 49 years and 50 to 69 years: comparison of relative and absolute benefit. Monogr Natl Cancer Inst 1997;22:79–86.
(*5*) Chu KC, Tarone RE, Kessler LG, et al. Recent trends in U.S. breast cancer incidence, survival and mortality rates. J Natl Cancer Inst 1996;88:1571–9.
(*6*) Dershaw DD, Abramson A, Kinne DW. Ductal carcinoma in situ: mammographic findings and clinical implications. Radiology 1989;170:411–5.
(*7*) Stomper PC, Connoly JL. Ductal carcinoma in situ of the breast: correlation between mammographic calcification and tumor subtype. AJR Am J Rotengenol 1992;159:483–5.
(*8*) Page DL, Dupont WD, Rogers LW, Landenberger M. Intraductal carcinoma of the breast: follow-up after biopsy only. Cancer 1982;49:751–8.
(*9*) Farrow JH. Current concepts in the detection of the earliest of the early breast cancers. Cancer 1970;25:468–77.
(*10*) Lagios MD, Margolin FR, Westdahl PR, Rose MR. Mammographically detected ductal carcinoma in situ. Frequency of local recurrence following tylectomy and prognostic effect of nuclear grade on local recurrence. Cancer 1989;63:618–24.
(*11*) Arnesson LG, Smeds S, Fagerberg G, Grontoft O. Follow-up of two treatment modalities for ductal cancer in situ of the breast. Br J Surg 1989;76:672–5.
(*12*) Hetelekidis S, Schnitt SJ, Morrow M, Harris JR. Management of ductal carcinoma in situ. CA Cancer J Clin 1995;45:244–53.
(*13*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.
(*14*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years [published erratum appears in Can Med Assoc J 1993;148:718]. Can Med Assoc J 1992;147:1459–76.
(*15*) Burhenne LJ, Burhenne HJ, Kan L. Quality-oriented mass mammography screening. Radiology 1995;194:185–8.
(*16*) Burhenne HJ, Burhenne LW, Goldberg F, Hislop TG, Worth AJ, Rebbeck PM, et al. Interval breast cancers in the screening mammography program of British Columbia: analysis and classification. AJR Am J Roentgenol 1994;162:1067–71.
(*17*) Kosary CL, Ries LAG, Miller BA, Hankey BF, Harras A, Edwards BK. SEER Cancer Statistics Review, 1973–1992: Tables and graphs. Bethesda (MD): National Cancer Institute; 1995: DHHS Publ No. (NIH)96–2789.
(*18*) Esserman L, Kerlikowske K. Should we recommend screening mammography for women aged 40 to 49 years? Oncology 1996;10:357–64.
(*19*) Harris JR, Lippman ME, Veronesi U, Willett W. Breast cancer (2). N Engl J Med 1992;327:390–8.
(*20*) Colditz GA, Willett WC, Hunter DJ, Stampfer MJ, Manson JE, Hennekens CH, et al. Family history, age, and risk of breast cancer. Prospective from the Nurses' Health Study [published erratum appears in JAMA 1993;270:1548]. JAMA 1993;270:338–43.
(*21*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276:33–8.
(*22*) Bird RE. Low-cost screening mammography: report on finances and review of 21,716 consecutive cases. Radiology 1989;171:87–90.
(*23*) Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases [published erratum appears in Radiology 1992;184:878]. Radiology 1992;184:39–43.
(*24*) Robertson CL. A private breast imaging practice: medical audit of 25,788 screening and 1,077 diagnostic examinations. Radiology 1993;187:75–9.
(*25*) Sienko DG, Hahn RA, Mills EM, Yoon-DeLong V, Ciesielski CA, Williamson GD, et al. Mammography use and outcomes in a community. The Greater Lansing Area Mammography Study. Cancer 1993;71:1801–9.
(*26*) Rosenberg RD, Lando JF, Hunt WC, Darling RR, Williamson MR, Linver MN. The New Mexico mammography project. Screening mammography performance in Albuquerque, New Mexico, 1991 to 1993. Cancer 1996;78:1731–9.
(*27*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.
(*28*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.
(*29*) Harris R, Leininger L. Clinical strategies for breast cancer screening: weighing and using the evidence. Ann Intern Med 1995;122:539–47.
(*30*) Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR Am J Roentgenol 1995;165:1373–7.
(*31*) Chang SW, Kerlikowske K, Napoles-Springer A, Posner SF, Sickles EA, Perez-Stable EJ. Racial differences in timeliness of follow-up after abnormal screening mammography. Cancer 1996;78:1395–402.
(*32*) Dixon J, Chetty U, Forrest A. Wound infection after breast biopsy. Br J Surg 1988;75:918–9.
(*33*) Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med 1991;114:657–61.
(*34*) Cockburn J, Staples M, Hurley SF, De Luise T. Psychological consequences of screening mammography. J Med Screen 1994;1:7–12.
(*35*) Ellman R, Angeli N, Christians A, Moss A, Chamberlain J, Macquire P. Psychiatric morbidity associated with screening for breast cancer. Br J Cancer 1989;60:781–4.
(*36*) Gram IT, Lund E, Slenker SE. Quality of life following a false positive mammogram. Br J Cancer 1990;62:1018–22.
(*37*) Elmore JG, Barton MB, Moceri VM, Fletcher SW. Cumulative risk of a false-positive mammogram over a 10-year period [abstract]. J Gen Intern Med 1997;12 Suppl:107.
(*38*) Fisher ER, Costantino J, Redmond C, et al. Response-blunting the counterpoint. Cancer 1993;75:1223–7.
(*39*) Page DL, Lagios MD. Pathologic analysis of the National Surgical Adjuvant Breast Project (NSABP) B-17 Trial. Unanswered questions remaining unanswered considering current concepts of ductal carcinoma in situ [editorial]. Cancer 1995;75:1219–22.
(*40*) Fisher B, Costantino J, Redmond C, Fisher E, Margolese R, Dimitrov N, et

al. Lumpectomy compared with lumpectomy and radiation therapy for the treatment of intraductal breast cancer. N Engl J Med 1993;328:1581–6.
(*41*) Schwartz GF, Finkel GC, Carcia JC, Patchefsky AS. Subclinical ductal carcinoma in situ of the breast. Treatment by local excision and surveillance alone. Cancer 1992;70:2468–74.
(*42*) Fisher ER, Costantino J, Fisher B, Palekar AS, Redmond C, Mamounas E. Pathologic findings from the National Surgical Adjuvant Breast Project (NSABP) protocol B-17. Intraductal carcinoma (ductal carcinoma in situ). The National Surgical Adjuvant Breast and Bowel Collaborating Investigators. Cancer 1995;75:1310–9.
(*43*) Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996; 275:913–8.
(*44*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Likelihood ratios for modern screening mammography. Risk of breast cancer based on age and mammographic interpretation. JAMA 1996;276:39–43.
(*45*) Black WC, Nease RF Jr, Tosteson AN. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. J Natl Cancer Inst 1995;87:720–31.
(*46*) Swanson MG, Ragheb NE, Lin CS, et al. Breast cancer among black and white women in the 1980s. Cancer 1993;72:788–98.
(*47*) White E, Lee CY, Kristal AR. Evaluation of the increase in breast cancer incidence in relation to mammography use. J Natl Cancer Inst 1990;82: 1546–52.
(*48*) Miller BA, Feuer EJ, Hankey BF. Recent incidence trends for breast cancer in women and the relevance of early detection: an update. CA Cancer J Clin 1993;43:27–41.
(*49*) Institute of Public Health. University of Cambridge, Department of Community Medicine, Cancer Intelligence Unit, 1997.
(*50*) Pauker SG, Kassirer JP. Contentious screening decisions: does the choice matter? [editorial]. N Engl J Med 1997;336:1243–4.

Note

This work was supported by an NCI-funded Breast Cancer SPORE grant, P50 CA58207, and an NCI-funded Breast Cancer Surveillance Consortium cooperative agreement, 1 U01 CA 63740.

Mammography Outcomes in a Practice Setting by Age: Prognostic Factors, Sensitivity, and Positive Biopsy Rate

*Michael N. Linver, Stuart B. Paster**

The separate unplanned analysis of women ages 40–49 in population-based randomized controlled trials has resulted in demonstration of statistically significant breast cancer mortality reduction due to screening mammography in only two of the individual trials, and in all such trials only through meta-analysis. Therefore, many researchers have utilized the surrogate endpoints of tumor size and axillary lymph node status to evaluate screening efficacy. For the present study, these endpoints were evaluated in an audit of 854 screen-detected cancers found in 147,125 mammographic examinations performed in women over 40 between 1988 and 1994 in a community practice setting. The concerns that mammography in the 40–49 group has a lower sensitivity and higher biopsy rate were also addressed. Median invasive tumor size and lymph node positivity were found to be equally small (1.0–1.1 cm and 13.5–12.2%, respectively), and the sensitivity and overall biopsy rate were found to be constant over all ages 40 and above. Positive biopsy rate (PBR) varied directly with increasing age, paralleling the measured cancer detection rate in each decade, with no abrupt change at age 50. We conclude that modern mammography in a community practice setting can successfully detect breast cancers with favorable prognostic factors and achieve constant sensitivity and acceptable PBRs in all women over 40. Our data also suggest that many of the large differences seen by inappropriately dividing data at age 50 decrease or disappear when analysis is performed by decade. [Monogr Natl Cancer Inst 1997;22:113–117]

The value of regular screening mammography in reducing breast cancer mortality for women age 50 and older has been generally accepted based on multiple randomized controlled trials (RCTs). However, the separate unplanned analysis of women ages 40–49 has caused confusion and disagreement over the benefit for these women, despite the fact that the results from these RCTs were not expected to permit this type of subgroup analysis. A mortality reduction has been seen in the 40–49 group in most RCTs, but because the RCTs were not designed to evaluate this age group exclusively, point estimates in most of the individual RCTs have not achieved statistical significance (*1,2*). It is only through longer follow-up and meta-analysis of these RCTs that a statistically significant difference in mortality between women invited to be screened and those not invited to screening with mammography has now been demonstrated (*1–3*). A single definitive randomized trial in the United States to test screening efficacy in the 40–49 group large enough to have the potential to achieve statistical significance would not only be difficult (at least 1.5 million women would need to be enrolled), but would not yield meaningful results for another 10–15 years (*4*). A trial requiring fewer women has been proposed in Europe but has not yet begun (*5,6*).

Several other issues have been raised. Some have suggested that comparing women ages 40–49 with all other women skews the analysis (*7*). It has also been suggested that the sensitivity of mammography is considerably lower among younger women (*8*) and that the biopsy rate is too high. Finally, the question of accuracy of mammography in the 40–49 age group in a community radiology practice setting remains, as very little recent data addressing this subject exist.

Given the difficulties with RCT analyses, many have suggested and employed surrogate endpoints to assess screening efficacy (*4,9–11*). We have undertaken an analysis in a community practice setting to assess these endpoints and address the other above-mentioned issues.

Methods

Our group, X-Ray Associates of New Mexico, is comprised of 12 general radiologists. All 12 radiologists interpret mammograms, as well as all other imaging modalities. Using four private outpatient offices, two community hospitals, and two mobile vans, we interpret approximately 30,000–40,000 mammograms yearly, 90% of which are screening studies. In 1988, we instituted a program to upgrade the quality of our mammography services. We all attended dedicated courses in mammography, upgraded our equipment and image quality, and undertook an extensive quality assurance program that included data collection and analysis. The benefits of these changes have previously been reported (*12*). These upgrades were virtually identical to the currently required minimum quality standards for mammography facilities throughout the United States as mandated by the federal Mammography Quality Standards Act (MQSA) of 1992, which became effective in October 1994 (*13*).

Through a computerized reporting system we designed and have utilized since 1988, we performed an audit of over 162,000

**Affiliation of authors:* X-Ray Associates of New Mexico, P.C., Albuquerque, NM.

Correspondence to: Michael N. Linver, M.D., X-Ray Associates of New Mexico, P.C., 715 Dr. Martin Luther King, Jr. Ave., N.E., Suite 112, Albuquerque, NM 87102.

See "Note" following "References."

mammograms performed on women over age 40 between February 1988 and December 1994. Of these, 147,125 were evaluated as screening mammograms (those performed on asymptomatic women). Approximately 25% of the screening mammograms in every patient-age decade beginning at age 40 were initial examinations, and 75% were subsequent examinations. This proportion did not vary by more than 2% in any decade. Diagnostic mammographic examinations were separated from the screening examinations on the basis of presenting breast pain, palpable lump, or nipple discharge.

The surrogate endpoints chosen to evaluate screening efficacy were tumor size and axillary lymph node status. These prognostic factors are the biological indicators that have been demonstrated to distinguish women whose prognosis is more favorable in the RCTs (*10*). In addition, because we were in the unique position at that time to compare our data with a statewide tumor registry via computer linkage, we were able to match 94% of our cases and evaluate the accuracy of mammography by tracking all false negatives and comparing resultant sensitivity values. We further evaluated mammography efficiency in detecting cancers by calculating positive biopsy rates (PBRs) [(biopsies positive for cancer)/(all biopsies performed based on mammographic recommendation for biopsy)]. We completed an evaluation for each age group, including an analysis by decade.

Results

We successfully diagnosed 1,303 cancers among women ages 40 and above. Approximately two thirds were screen-detected (Table 1). When only the screen-detected cancers were evaluated, the cancer detection rate was 3.6 per 1,000 screening cases for women ages 40–49, 4.8 per 1000 for ages 50–59, 6.9 per 1,000 for ages 60–69, 9.5 per 1000 for ages 70–79, and 12.4 per 1,000 for women over 80 (Table 2). As would be expected, these rates reflect the prior probability of breast cancer and are proportional to the incidence expected in these age groups based on Surveillance, Epidemiology, and End Results (SEER) data (*14*). If the data are grouped such that women ages 40–49 are compared to all those 50 and over, the proportion of cancers diagnosed in these two groups is grossly unbalanced (171 to 683), and the cancer detection rate appears vastly different (3.6 per 1,000 for the 40–49 group and 6.8 per 1,000 for the over 50 group) (Table 1). The evaluation by decade (Table 2), however, shows a more gradual incremental increase consistent with the prior probability of cancer in each decade.

Table 1. Cancers detected

	Ages	
	40–49 group	Over 50 group
Total mammographic exams	53,583	109,023
Screening mammograms	47,561	99,564
All cancers found mammographically	265	1,038
Asymptomatic (screen-detected) cancers found mammographically	171	683
Cancer detection rate*	3.6	6.8

*Number of cancers detected per 1,000 screening examinations.

Review of prognostic factors for screen-detected cancers showed that 78% in the 40–49 group were minimal cancers (ductal carcinoma *in situ* [DCIS] or invasive cancers 1 cm or less), compared with 61% in women over 50 (Table 3). DCIS comprised 41% of cancers in the 40–49 group, as compared with 32% in the 50–59 group, 17% in the 60–69 group, 15% in the 70–79 group, and 14% in the over 80 group (Table 4). Again, comparing the 40–49 group to all those over age 50 shows a markedly disparate rate of 41% to 20%, but when comparing rates by decade, the difference between the DCIS rate in the 40–49 group and that in the 50–59 group is considerably smaller (41% to 32%). The DCIS detection rate was found to be relatively constant (1.2 to 1.8 cases per 1,000 screening exams) in all decades (Table 4).

Median size of the invasive cancers was 1.0 cm in the 40–49 group and 1.1 cm in the over 50 group (Table 3). Axillary lymph node positivity was 13.5% in the 40–49 group and 12.2% in the over 50 group (Table 3).

Axillary lymph node status of small screen-detected invasive cancers 1 cm or less in size yielded similarly low node positivity values of 8% for the 40–49 group and 7% for the over 50 group (Table 5).

Using the "one year" definition of false negative (cancer detected within one year of "negative" screening), we calculated similar sensitivity values of 86.8% for the 40–49 group and 87.2% for the over 50 group (Table 6).

Overall biopsy rates on screen-detected abnormalities were virtually the same in all decades, varying from 1.44% in the 40–49 group to 1.78% in the over 80 group (Table 7).

PBRs were 25% in the 40–49 group, 32% for the 50–59 group, 41% for the 60–69 group, 60% for the 70–79 group, and 70% for the group over 80 (Table 7). This is clearly a steady, gradual change. However, if women ages 40–49 are compared to all women over 50, the 25% PBR seems much lower than the 43% found in the group over 50 (Table 6).

Discussion

Day and others have argued persuasively that intermediate measures are useful for the evaluation of a screening program, serving as proxies for conventional endpoints such as death from breast cancer (*10,11*). The surrogate endpoints chosen here for evaluating screening efficacy—tumor size and axillary lymph node positivity—have both been shown to correlate inversely with survival (*15,16*): when tumor size is small and axillary lymph node metastasis is absent, survival in all age groups over 40 is much greater. Our findings reflect favorable measures for both parameters in women 40–49 (Table 3) and show no significant differences in any age group over 40, paralleling results in other recent studies (*15–18*). These data would imply that, as demonstrated by Tabár (*15*) and Therfjell and Lindgren (*16*), all women over 40 have an equally high likelihood of long survival, when small, node-negative tumors are detected by screening. These same prognostic indicators correlate well with mortality results found in the RCTs.

When evaluating screen-detected invasive cancers 1 cm or smaller, we found an even more impressive prognostic indicator in the lower axillary lymph node positivity (7–8%) in all women

Table 2. Cancer detection rates for screening cases, by decade

	Ages				
	40–49	50–59	60–69	70–79	Over 80
Screening mammograms	47,561	40,005	34,402	20,675	4,482
Screen-detected cancers found mammographically	171	192	239	196	56
Cancer detection rate*	3.6	4.8	6.9	9.5	12.4

*Number of cancers detected per 1,000 screening examinations.

Table 3. Screen-detected cancers: size and nodal status

	Ages	
	40–49 group	Over 50 group
Screening-detected cancers, total	171	683
DCIS	70 (41%)	139 (20%)
Invasive cancers	101 (59%)	544 (80%)
Minimal cancers (DCIS or ≤1 cm)	78%	61%
Median size (invasive cancers only)	1.0 cm	1.1 cm
Axillary lymph node positivity	13.5%	12.2%

over 40 (Table 4). Our findings were similar to those of Curpen, Sickles *et al.* (*17*), supporting the hypothesis that advancing the time of diagnosis at any age reduces the likelihood for axillary lymph node metastasis, thus improving prognosis.

One could anticipate that this evidence would further translate into a reduction in mortality for all women screened at age 40 and older, although the many biases intrinsic in the use of survival data warrant caution (*9*). Nevertheless, long-term survival provides confidence that a benchmark of cure has been achieved. The fact that our findings also parallel those reported in academic institutions would seem to support their reproducibility outside the academic setting (*17*).

Further, our detection of smaller, node-negative breast cancers was accomplished with a high degree of accuracy, regardless of patient age: a sensitivity in the 86–87% range was achieved in all women over 40 (Table 5). This finding would appear to refute the contention made by some that the value of screening under age 50 is compromised by markedly lower sensitivity (*8*). Our data strongly suggest that mammography in women 40–49 has an equally high likelihood of finding tumors with a favorable prognosis as in women 50 and over. The lack of significant difference in sensitivity here, as contrasted to the sizable differences in sensitivity seen in many of the earlier RCTs, which showed much lower sensitivity on the 40–49 group, may be explained in part by the advances in imaging technology that have occurred since these earlier studies (and that are now mandated by the MQSA), especially regarding imaging of the dense breast pattern more often seen in younger women (*19,20*).

The overall biopsy rate in each decade maintained a virtually constant value of 1.44–1.78%, with the lowest rate in the 40–49 group. This finding runs counter to the notion that a higher biopsy rate is an automatic negative feature of screening women ages 40–49.

The PBR varied directly with increasing age, with no abrupt change at age 50. This merely reflected the prior probability of cancer, as demonstrated by the increasing cancer detection rate we found with increasing age (Table 7). When curves for PBR and cancer detection rate were plotted by decade using our data, the two curves were seen to run virtually in parallel (Fig. 1). A major change could be made to appear at age 50 by grouping women 40–49 and comparing them to all women over age 50, but this was artificial. Indeed, the often dramatic differences in prognostic indicators and other screening data between the 40–49 group and those over 50 cited by others (*8*) may well be explained by the conventional arbitrary division of data at age 50, creating two comparison age groups of widely unequal duration. The example described here involving PBR illustrates this point well: the PBR did not hit a statistical wall at age 50 and suddenly jump from 25% to 43%, as analysis of only the 40–49 group and the over 50 group would intimate. By grouping the same data by decade, we enabled the true pattern of steady, gradual increase (from 25%, to 32%, to 41%, etc.) in each ensuing decade to emerge.

Certainly, the greater number of benign biopsies in the 40–49 group could well be construed as a risk of screening, but the same could be said of the 50–59 group in our analysis relative to those in their sixties, of women in their sixties relative to those

Table 4. Screen-detected cancers: invasive cancers and DCIS, by decade

	Ages					
	40–49	50–59	60–69	70–79	80+	Total
Screening mammograms	47,561	40,005	34,402	20,675	4,482	147,125
Total cancers	171	192	239	196	56	854
Invasive cancers	101 (59%)	131 (68%)	198 (83%)	167 (85%)	48 (86%)	645 (76%)
DCIS	70 (41%)	61 (32%)	41 (17%)	29 (15%)	8 (14%)	209 (24%)
DCIS detection rate*	1.5	1.5	1.2	1.4	1.8	1.4

*Number DCIS cases detected per 1,000 screening examinations.

Table 5. Axillary lymph node positivity: invasive cancers ≤1 cm (screen-detected cancers only)

	Ages	
	40–49 group	Over 50 group
Cancers <1 cm	51	226
Positive axillary lymph nodes	4 (8%)	15 (7%)

Table 6. Sensitivity* and positive biopsy rate for screening cases

	Ages	
	40–49 group	Over 50 group
Screening mammograms	47,561	99,564
Biopsies done based on mammographic findings	684	1,590
Cancers found at biopsy (and correctly identified mammographically)	171	683
False negative cases†	26	100
Overall sensitivity	86.8%	87.2%
Positive biopsy rate	25%	43%

*Sensitivity: (number of true positives)/(number of true positives + number of false negatives) × 100.

†False negative: detection of cancer within one year of mammographic examination with normal findings.

in their seventies, and so forth. Nonetheless, all groups demonstrated PBRs in the acceptable range of target values endorsed in the Agency for Health Care Policy and Research (AHCPR) Clinical Practice Guidelines of Quality Determinants of Mammography (*21*). We therefore believe PBRs in these ranges are justified in our practice and within our community, especially in view of the high rate of small, node-negative tumors we detected through screening.

Note is made that, while the DCIS detection rate was constant across all decades, a much higher ratio of DCIS to invasive cancer was found in the 40–49 and the 50–59 age groups, as compared to the 60 and over groups, which demonstrated almost identical lower ratios in each decade over 60. Whether this difference reflects the change in cancer from primarily intraductal to invasive disease as women age, a fundamentally different kind of cancer manifesting itself in younger women, or a greater detection of indolent DCIS cases in the early rounds of screening as each cohort passes from one decade to the next remains a major point for future research.

Future research should also encourage others in academia and community practice to perform and publish similar audits of their screening mammography practices. It is clear that the quality of mammography practiced throughout the United States has improved dramatically with the implementation of the MQSA, as shown in recent studies (*22*). This improvement in quality could be measured quantitatively if widespread outcomes audits

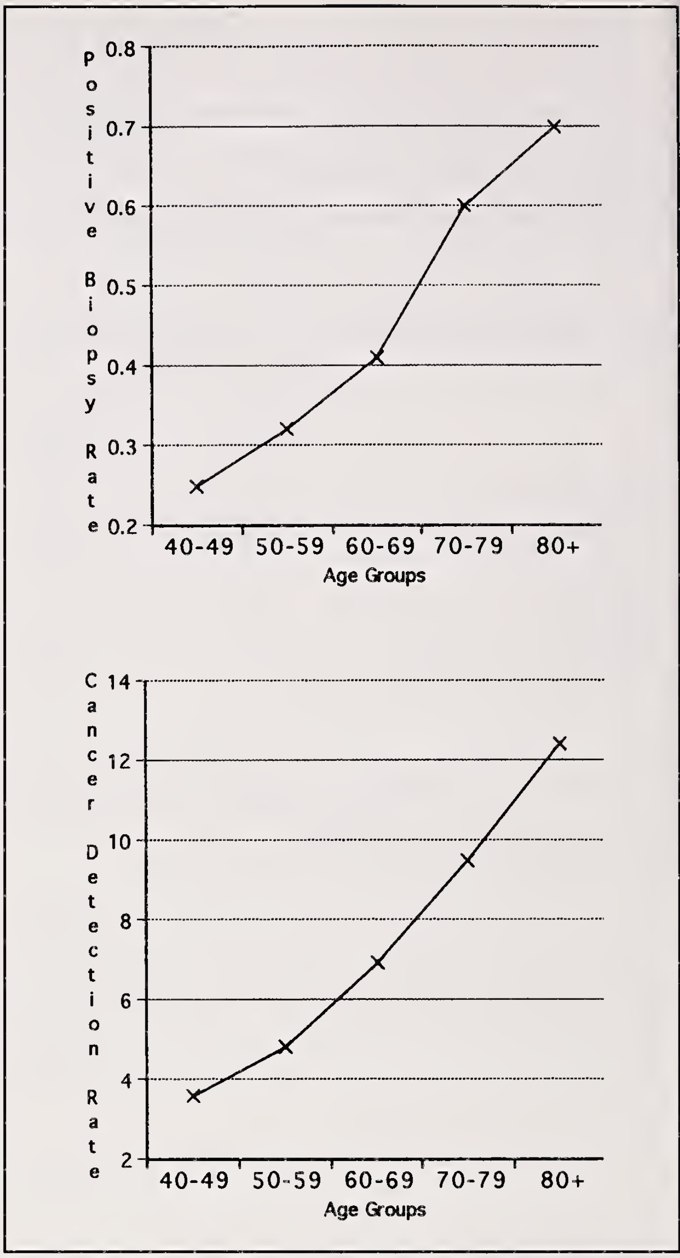

Fig. 1. Positive biopsy rate and cancer detection rate for screening cases: comparison by decade.

Table 7. Overall biopsy rate and positive biopsy rate for screening cases, by decade

	Ages				
	49–49	50–59	60–69	70–79	Over 80
Screening mammograms	47,561	40,005	34,402	20,675	4,482
Biopsies done based on mammographic findings	684	600	583	327	80
Cancers found at biopsy (and correctly identified mammographically)	171	192	239	196	56
Overall biopsy rate	1.44%	1.50%	1.69%	1.58%	1.78%
Positive biopsy rate	25%	32%	41%	60%	70%

were performed and published. However, this has not been the case, primarily due to the lack of protection of audit data from medical-legal discovery in most states (*23*). Passage of national legislation to protect audit data is needed to permit a more accurate overall evaluation of the performance of mammography in women 40 and over in the community setting.

Conclusion

We find modern screening mammography in the community practice setting to be as successful in detecting breast cancers with favorable prognostic factors in women age 40–49 as in women over 50. These results were attained through the early application of the high-quality standards of modern screening mammography that are now mandated by federal law. Our findings corroborate the recent favorable results of others who have similarly evaluated screening efficacy via surrogate endpoints (*15–18*). We also find evidence to suggest that advancing the time of diagnosis at any age reduces the likelihood of axillary lymph node metastasis, thereby improving prognosis. Further, we find the sensitivity of mammography to be constant as a function of age. These results are attainable without an unacceptably large number of biopsies in any decade. Although the PBR is lowest in the 40–49 decade, it parallels our cancer detection rate by decade, reflecting the prior probability of cancer by age. As with the other measured prognostic factors and with sensitivity, we find PBR to show no dramatic change at age 50. Rather, our data suggest that the apparent large differences in screening outcomes seen by inappropriately dividing and comparing groups at age 50 are created artificially and decrease or disappear when grouping is performed by decade.

References

(*1*) Smart CR, Hendrick RE, Rutledge JH III, Smith RA. Benefit of mammography screening in women ages 40 to 49 years. Current evidence from randomized controlled trials [published erratum appears in Cancer 1995; 75:2788]. Cancer 1995;75:1619–26.

(*2*) Bjurstam N, Bjorneld L, Duffy SW. The Gothenburg Breast Screening Trial: results from 11 years follow-up. In: NIH Consensus Development Conference, Breast Cancer Screening for Women Ages 40–49, Program and Abstracts. Bethesda (MD): National Institutes of Health, 1997:63–4.

(*3*) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer 1996;68:693–9.

(*4*) Eckhardt S, Badellino F, Murphy GP. UICC meeting on breast cancer screening in pre-menopausal women in developed countries. Geneva, 29 September–1 October 1993. Int J Cancer 1994;56:1–5.

(*5*) Rosselli del Turco M. Eurotrial 40: A randomized population-based trial on the efficacy of mammographic screening for breast cancer in pre-menopausal women [abstract]. Presented at Board of Directors Meeting, American Cancer Society, Atlanta, GA, March 1997.

(*6*) Smith RA, personal eommunication, 1997.

(*7*) Kopans DB, Halpern E, Hulka CA. Statistical power in breast cancer screening trials and mortality reduction among women 40–49 years of age with particular emphasis on the National Breast Screening Study of Canada. Cancer 1994;74:1196–203.

(*8*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276:33–8.

(*9*) Feig, SA. Determination of mammographic screening intervals with surrogate measures for women aged 40–49 years [editorial]. Radiology 1994; 193:311–4.

(*10*) Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.

(*11*) Day NE. Quantitative approach to the evaluation of screening programs. World J Surg 1989;13:3–8.

(*12*) Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases [published erratum appears in Radiology 1992;184:878]. Radiology 1992;184:39–43.

(*13*) PL 102-539. The mammography quality standards act of 1992, paragraph 354.

(*14*) American Cancer Society. Cancer Facts and Figures-1995. Atlanta (GA): American Cancer Society; 1995.

(*15*) Tabar L, Duffy SW, Burhenne LW. New Swedish breast cancer detection results for women aged 40–49. Cancer 1993;72(4 Suppl):1437–48.

(*16*) Therfjell EL, Lindgren JA. Breast cancer survival rates in women younger and in those older than 50 years: effect of mammography screening. Radiology 1996;201:421–6.

(*17*) Curpen BN, Sickles EA, Solitto RA, Ominsky SH, Galvin HB, Frankel SD. The comparative value of mammographic screening for women 40–49 years old versus 50–64 years old. AJR Am J Roentgenol 1995;164: 1099–103.

(*18*) Burhenne LJ, Burhenne HJ, Kan L. Quality-oriented mass mammography screening. Radiology 1995;194:185–8.

(*19*) Young KC, Wallis AG, Ramsdale ML. Mammographic film density and detection of small breast cancers. Clin Radiology 1994;49:461–5.

(*20*) Jackson VP, Hendrick RE, Feig SA, Kopans DB. Imaging of the radiographically dense breast. Radiology 1993;188:297–301.

(*21*) Bassett LW, Hendrick RE, Bassford TL, et al. Quality determinants of mammography: clinical practice guideline no. 13. AHCPR Publ No. 95-0632. Rockville (MD): DHHS, PHS, AHCPR, 1994.

(*22*) Nadel MV. Mammography services: initial impact of new federal law has been positive: report no. GAO/HEHS-96-17. Washington (DC): US General Accounting Office Report to Congressional Committees, Health Education and Human Services Division, 1995.

(*23*) Linver MN, Rosenberg RD, Smith RA. Mammography outcomes analysis: potential panacea or Pandora's box? [commentary]. AJR Am J Roentgenol 1996;167:373–5.

Note

We thank Robert A. Smith, Ph.D., Daniel B. Kopans, M.D., R. Edward Hendrick, Ph.D., Carl J. D'orsi, M.D., László Tabár, M.D., and Robert D. Rosenberg, M.D., for their help in preparing the manuscript.

Radiation Risk From Screening Mammography of Women Aged 40–49 Years

*Stephen A. Feig, R. Edward Hendrick**

Although direct evidence of carcinogenic risk from mammography is lacking, there is a hypothetical risk from screening because excess breast cancers have been demonstrated in women receiving doses of 0.25–20 Gy. These high-level exposures to the breast occurred from the 1930s to the 1950s due to atomic bomb radiation, multiple chest fluoroscopies, and radiation therapy treatments for benign disease. Using a risk estimate provided by the Biological Effects of Ionizing Radiation (BEIR) V Report of the National Academy of Sciences and a mean breast glandular dose of 4 mGy from a two-view per breast bilateral mammogram, one can estimate that annual mammography of 100,000 women for 10 consecutive years beginning at age 40 will result in at most eight breast cancer deaths during their lifetime. On the other hand, researchers have shown a 24% mortality reduction from biennial screening of women in this age group; this will result in a benefit-to-risk ratio of 48.5 lives saved per life lost and 121.3 years of life saved per year of life lost. An assumed mortality reduction of 36% from annual screening would result in 36.5 lives saved per life lost and 91.3 years of life saved per year of life lost. Thus, the theoretical radiation risk from screening mammography is extremely small compared with the established benefit from this life-saving procedure and should not unduly distract women under age 50 who are considering screening. [Monogr Natl Cancer Inst 1997;22: 119–124]

The risk of radiation-induced breast cancer is a consideration in determining the advisability of mammographic screening for women of any age group and may be especially important for women aged 40–49 years. Due to the relatively lower breast cancer incidence in younger women, it is particularly important to assess in these women the number of lives saved versus deaths caused and the years of life expectancy gained per year of life lost through screening.

Risk Assessment

Although no women have ever been shown to have developed breast cancer as a result of mammography, not even from multiple examinations received over many years at mean glandular doses considerably higher than the current average mammographic doses of 3–4 mGy (0.3–0.4 rad), the possibility of such risk exists because excess breast cancers have been observed among populations receiving much higher doses—say, 0.25–20 Gy (25–2,000 rads). These include Japanese A-bomb survivors (*1*), North American tuberculosis sanitoria patients from Massachusetts (*2*) and Canada (*3*) who underwent multiple chest fluoroscopies, women from New York State (*4*) and Sweden (*5*) treated with radiation therapy for benign breast conditions such as postpartum mastitis, and women who had been treated in California with radiation therapy for Hodgkin's disease (*6*).

Estimating the risk of breast cancer from low-dose radiation is complex. However, relatively similar estimates have been made by various committees over the past 20 years, most notably by the 1977 National Cancer Institute (NCI) Ad Hoc Working Group on the risks associated with mammography and mass screening for the detection of breast cancer (*7*), by the 1980 Committee on the Biological Effects of Ionizing Radiation (BEIR III) of the National Academy of Sciences (*8*), by the 1985 National Institutes of Health Ad Hoc Group to Develop Radiolepidemiological Tables (*9*), by the National Academy of Sciences' 1990 National Research Council Committee on the Biological Effects of Ionizing Radiation (BEIR V) (*10*), and by the 1994 United Nations Scientific Committee on the Effects of Atomic Radiation (*11*). Each committee has had to base its estimate not only on the follow-up data available at that time, but also on a selection of other assessment options, such as dose-response models, length of latent period, duration of radiation effect, age-related radiation sensitivity, and absolute versus relative risk models.

Dose-Response Models

Because radiation-induced and spontaneously occurring breast cancers cannot be distinguished histologically (*12,13*), the presence of radiation-induced tumors can only be established statistically if a significant number of excess cancers are observed in an exposed population. This type of inference becomes harder and harder to establish as lower and lower doses are considered, since the number of exposed women required to demonstrate an effect is related to the inverse square of dose. For example, if 1,000 exposed and 1,000 control women are needed to demonstrate an effect at 1 Gy, then two groups of 100,000 women each are necessary at 0.1 Gy and two groups of 10,000,000 women each are necessary at 1 cGy, assuming a linear dose-response relationship (*14*).

If there is any risk to the breast from doses in the mammographic range (3–4 mGy per two-view exam) or even from doses

**Affiliations of authors:* S. A. Feig, Jefferson Medical College, Philadelphia, PA; R. E. Hendrick, University of Colorado, Health Sciences Center, Denver, CO.

Correspondence to: Stephen A. Feig, M.D., Breast Imaging Center, Thomas Jefferson University Hospital, 1100 Walnut Street, Philadelphia, PA 19107–5563.

of 100 mGy (10 rad) or less, the magnitude of the risk may be estimated by means of dose-response curves, which describe the possible relationship between radiation dose and radiogenic cancer incidence (Fig. 1). In the linear dose-response model, incidence is directly proportional to dose: if the dose is diminished by a factor of 10, the excess cancer incidence will also be reduced by the same factor. With the quadratic dose-response relationship, the effect is proportional to the dose squared: if the dose is reduced by a factor of 10, the number of excess cancers would be reduced by a factor of 100. The linear-quadratic dose-response relationship predicts a risk between the risks expected from pure linear and pure quadratic models.

Most but not all experiments on a wide variety of radiation-induced tumors in laboratory animals exhibit a quadratic dose-response relationship at doses below 0.5 Gy (50 rads) (*10*). However, a similar relationship may not necessarily hold for breast cancer in humans.

Most studies on radiation-induced breast cancer in humans contain a paucity of data on doses below 0.5 Gy (50 rads), and not one provides direct information concerning risks from doses less than 0.1 Gy (10 rads) (*15*). However, results from a linear regression analysis over a wide range of doses found data highly consistent with a linear model; the data also fit a linear-quadratic model fairly well when a strong linear component is present (*1*). Nevertheless, a quadratic dose-response function at doses below 0.05 Gy (5 rads) cannot be excluded at the 95% confidence level (*1*). Therefore, the linear model is most often used to estimate risk at low doses. Lower risk estimates would be obtained with other types of dose-response relationships. Although an appropriate upper confidence limit of a linear coefficient represents the upper limit of risk, a point estimate of the slope of a linear fit provides a reasonable estimate of risk.

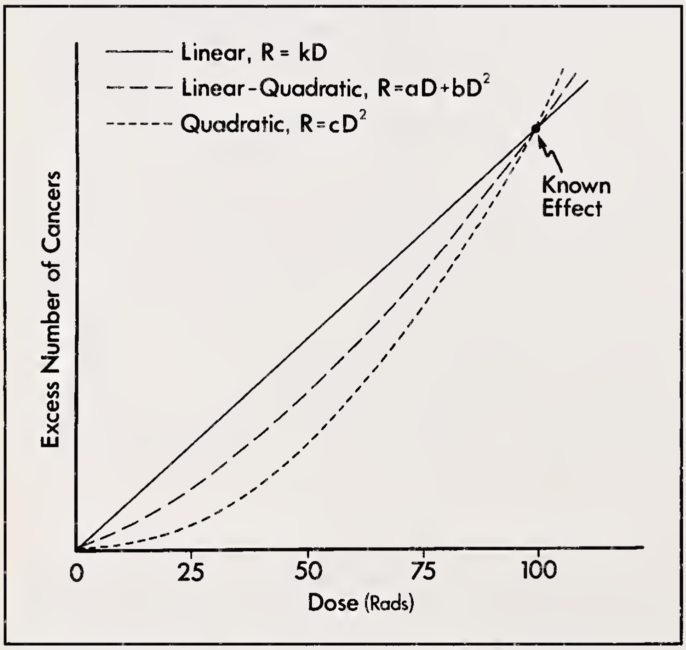

Fig. 1. Models for possible dose-response relationships at low doses. Most estimates for the hypothetical breast cancer risk from mammography have employed a linear dose-response model with the understanding that this projection represents the upper limits of such risk. R = risk per rad.

Latent Period and Duration

The latent period refers to the minimal length of time between exposure and earliest demonstration of excess cancers in a population. Because radiogenic breast cancers do not occur earlier than the spontaneous variety, the latent period may depend on age at exposure. Most reports have assumed latent periods of at least 10 years and a lifetime persistence of radiation risk in the exposed population. The BEIR V Report assumed that there is a latent period of about 10 years after exposure before the risk of radiation-induced breast cancer is non-negligible. The Report also assumed that the period of excess risk may persist for the patient's lifetime, since all populations have continued to exhibit excess breast cancer risk on the longest follow-up studies—those following subjects 30–45 years after exposure (*1–4*).

Age at Exposure

All but one of the studies of radiogenic risk found decreased risk with increasing age at exposure (*1–3,5,6*) (Fig. 2). New York women treated with radiotherapy for postpartum mastitis (*4*) constitute the only group that has not shown any relationship between risk and age at exposure. Their breasts were, however, in a proliferative state, with elevated hormonal stimulation due to parturition and lactation. The BEIR V Report concluded that "there is little evidence of any increased risk to women exposed after age 40" (*10*).

Additive and Relative Risk Models

Additive and relative risk models represent two different ways of estimating excess risk (defined as either excess breast cancer incidence or mortality) following radiation. Additive (or absolute) risk estimates are given as a number of excess cancers/million women/year/cGy (rad). Relative risk estimates are given as the percentage increase in the natural breast cancer incidence/year/cGy (rad). BEIR V used a time-dependent relative risk model in which relative risk varied over time during the follow-up, reaching a peak at 15–20 years after exposure and then declining (*10*). Recent studies suggest that the complexity of BEIR V model may not be necessary to explain these data (*1*). BEIR V used the relative risks derived from mortality data from the Japanese and non-Nova Scotia Canadian populations to provide an absolute risk estimate for mortality among North American women according to age at exposure. Although the excess relative risk for Japanese women was 2 to 3 times that for non-Nova Scotia Canadian women, this difference was not statistically significant ($P = 0.12$). Although Japanese background breast cancer rates are considerably lower than those in Canada, the additive excess risk per unit dose was not significantly less than that for non-Nova Scotia women ($P>.5$) (*10*).

Quantifying Benefits and Risks

Using the 1985 NIH relative risk estimate, Feig and Ehrlich found that a single screen of women at ages 40–44 and 45–49 with a dose of 2.5 mGy and 20% reduction in breast cancer mortality due to screening would result in benefit/risk ratios of 35 and 90 years of life expectancy gained per year of life lost respectively (*15*).

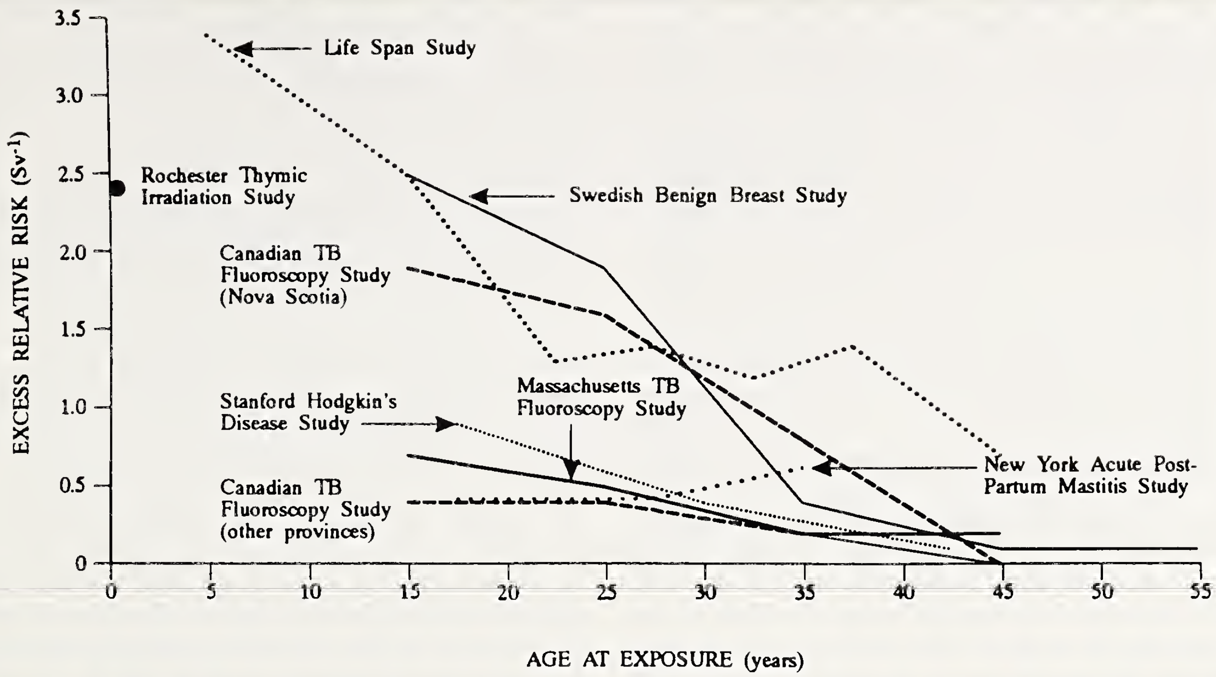

Fig. 2. Excess relative risk per 0.1 Sievert (0.1 Gy absorbed dose) for breast cancer incidence according to age at exposure. From reference *(11)* with permission.

Using the 1990 BEIR V relative risk estimate, Feig et al. (*16*) calculated that a single mammographic screening of women at age 45 with a dose of 2.5 mGy and breast cancer mortality reductions of 20% and 40% due to screening would avert 30 and 60 deaths per death caused respectively. Assuming that some radiogenic cancers would be detected by subsequent screening, the benefit/risk ratios from the single screen would be 37.5 and 100 respectively at the same levels of benefit.

Law calculated that a single mammographic film per breast with a dose of 1 mGy at age 40–49 would detect 186 times more breast cancers than it might induce (*17*).

Based on the 1994 Radiation Effects Research Foundation (RERF) relative risk estimate, Mettler et al. developed benefit/risk ratio tables comparing fatal cases of breast cancer prevented by screening mammography to those caused by screening mammography (*18*). Mortality reductions of 15% for screening women age 40–49 and 25% for screening women age 50–75 were assumed along with a dose of 2.8 mGy per two-view mammographic examination. The authors calculated that if a woman began annual mammography at age 40, mammographic examination at age 44 would provide 850 times more benefit than the potential harm from all of her mammographic examinations combined.

Current Estimates of Screening Benefit

More accurate quantitative information on reduction in breast cancer screening mortality through screening has become available during the past several years through longer-term follow-up of women enrolled in randomized controlled trials (RCTs). Two separate meta-analyses of data from seven population-based RCTs have both shown a breast cancer mortality reduction of about 24% from screening women aged 40–49 years at entry in intervals of generally every two years (range = 12–28 months) (*19,20*). Specifically, a relative mortality reduction of 0.76 (95% confidence interval [CI]: 0.61–0.98) was found by Smart et al. (*19*), and a reduction of 0.76 (95% CI: 0.62–0.93) was found by the Organizing Committee, Falun Sweden Screening Meeting (*20*). For women aged 50 and over invited for biennial screening in the Swedish Two-County Trial, a statistically significant 39% reduction in breast cancer mortality has been observed (*20*).

Based on relative death hazards found for cancers detected at screening, for interval cancers, for cancers found among study group women who refused to be screened, and for those among control group women, it has been calculated that if all study group women in the two-county trial had been screened every year, a breast cancer mortality reduction of 36% could be expected for those aged 40–49 years at entry (*20,21*), and a 45% mortality reduction in breast cancer mortality could be expected for those aged 50–74 years at entry (*20*).

Current Radiation Risk Estimates

Recently, it has been suggested that the mean glandular dose for a two-view per breast mammographic examination could be 3–4 mGy higher than the previous estimate of 2.5 mGy. This higher estimate is due to a larger estimated compressed breast thickness (5–5.7 cm vs. 4.2 cm) (*17,22*) and increased x-ray exposures to attain higher average optical densities (1.4–1.8 vs. 1.3) on the mammographic film. Higher optical densities have been shown to result in earlier detection of breast cancer (*23*).

The BEIR V Report estimated mortality from radiation-induced cancers based on a combined analysis of data from Japanese atomic bomb survivors and non-Nova Scotia Canadian

tuberculosis patients receiving multiple chest fluoroscopies. Using an age-at-exposure–dependent and time-since-exposure–dependent relative risk model, a linear dose-response relationship, and a 10-year latent period, the BEIR V Committee estimated that if 100,000 U.S. women aged 40–49 years received a single dose of 10 rem (100 mGy), at worst no more than 20 excess breast cancer deaths might occur during the lifetimes of those 100,000 women.

Based on this estimate, it ean be calculated that if 100,000 women were to receive annual mammography for 10 consecutive years beginning at age 40 with a dose of 4 mGy per examination, at most 8 breast cancer deaths might result over the lifetimes of these 100,000 women. However, if these women continued to be screened after age 50, some radiation-induced breast cancers would be detected at a curable stage at a subsequent screen. Assuming mortality reductions of 39% for biennial screening and 45% for annual screening of women age 50 and over, one can estimate the number of breast cancer deaths potentially caused by annual screening of 100,000 women in their forties to be 4.9 deaths and 4.4 deaths respectively (Table 1).

On the other hand, 5 biennial screenings of 100,000 women beginning at age 40 might at worst result in 4 excess breast cancer deaths. Subsequent biennial or annual screening beginning at age 50 would reduce the number of deaths from breast cancers potentially induced by screening 100,000 women age 40–49 to 2.4 deaths and 2.2 deaths respectively (Table 1).

Benefit/Risk Ratio Expressed as Lives Saved per Life Lost

Deaths averted through screening women in their forties can be calculated among 100,000 women aged 40–49 years; a natural breast cancer incidence at 1,620 invasive breast cancers/year can be expected over the 10-year period between each woman's 40th and 50th birthdays based on the National Cancer Institute Surveillance, Epidemiology, and End Results Program (SEER) data (*24*). Assuming a 20-year relative survival rate of 50% for these invasive cancers in the absence of screening (*24*), one can expect at least 810 breast cancer deaths due to these breast cancers. At the same time, biennial sereening—shown to produce a 24% mortality reduction (*19,20*)—could prevent 194 of these breast cancer deaths. Likewise, assuming a 36% mortality reduction from annual screening (*20,21*), one can estimate that 292 of these breast cancer deaths would be prevented.

Therefore, annual screening of women age 40–49 years could save 36.5 (292/8) lives for every life potentially lost due to radiation-induced breast cancer, and biennial screening could save 48.5 (194/4) lives for every life potentially lost due to a radiation-induced cancer (Table 1). This is a fairly conservative estimate, since it assumes that no radiation-induced cancers are detected at a eurable stage due to screening subsequent to age 49. Subsequent biennial screening after age 50 could result in an improved benefit/risk ratio, and annual screening after age 50 would result in an even higher benefit/risk ratio for lives saved per life lost due to screening women age 40–49. If annual screening after age 50 were to reduce breast cancer deaths by 45%, benefit/risk ratios from screening women in their forties would be nearly twice as high as without screening after age 49. Given the current screening practice in the U.S., it is unlikely that a woman who went for annual or biennial screening during her forties would suddenly stop being screened after age 50. Therefore, most realistic benefit/risk ratios for women undergoing annual screening in their forties would range from 60/1–66/1 lives saved per life lost. For women undergoing biennial screening in their forties, the range would be from 81/1 to 88/1 lives saved per life lost (Table 1).

Table 1. Benefit/risk ratio expressed as lives saved due to mammographic screening of women aged 40–49 years* versus lives lost due to possible risk from radiation†

Screening interval	Screening after age 50		
	None	Biennial	Annual
Annual	36.5 (292/8)	59.6 (292/4.9)	66.4 (292/4.4)
Biennial	48.5 (194/4)	80.8 (194/2.4)	88.2 (194/2.2)

*Benefit estimate based on an average annual breast cancer incidence, a 20-year survival rate from SEER data (*24*), a 36% mortality reduction expected from annual screening (*20,21*), and a 24% mortality reduction observed from generally biennial screening in population-based randomized trials (*19,20*). Biennial and annual screening after age 50 is assumed to reduce deaths from radiation-induced breast cancer by 39% and 45%, respectively (based on data from reference *20*).

†Risk estimate based on BEIR V Report (*10*) and a mean glandular dose of 4 mGy per two-view/breast bilateral mammogram.

Benefit/Risk Ratio Expressed as Years of Life Expectancy Saved/Lost

Benefits and risks may also be compared as years of life gained through screening versus years of life potentially lost due to radiation-induced cancers. This can be better understood by means of the following calculations. Since nearly all deaths from breast cancer will occur within 20 years of diagnosis, the average death from breast cancer, whether naturally occurring or radiation induced, will occur around 10 years from diagnosis. Aecording to BEIR V, no radiation-induced breast cancer will occur within 10 years of radiation exposure, and the most likely time of detection of radiation-induced breast cancers will be 15 years after exposure. Since the average age at death occurs 10 years after detection, the average age at death from radiation-induced breast cancers due to screening women ages 40–49 years will be around age 70. Since the normal life span is 80 years, a woman who dies from breast cancer induced by screening during her forties will have lost an average of 10 years of life expectancy. On the other hand, the average age of death from breast cancer occurring between age 40–49 years would be age 55 or perhaps slightly older. Therefore, the average life saved through screening women aged 40–49 will add around 25 years of life expectancy. The ratio of the number of years of life expectancy saved versus lost through screening women in their forties will be 2.5 (25/10) times the ratio of lives saved versus lost from screening women in this age group (Table 2).

Assuming no further screening after age 49 and a 36% mortality reduction from annual screening, women age 40–49 will gain 91.3 years of life expectancy for every year possibly lost from radiation-induced cancers. For biennial screening, there will be 121.3 years of life expectancy gained per year potentially lost. As previously discussed, it is realistic to assume that women will continue to be screened every year or two after age

Table 2. Benefit/risk ratio expressed as years of life saved due to mammographic screening of women aged 40–49 years versus years of life lost due to possible risk from radiation*

	Screening after age 50		
Screening interval	None	Biennial	Annual
Annual	91.3	149.7	166.0
Biennial	121.3	198.9	220.5

*For mammographic screening of women aged 40–49, years of life expectancy gained/lost are 2.5 × lives saved/lost (see text for calculation). Lives saved/lost as per Table 1.

50, so that some radiation-induced cancers will be detected at a curable stage. In that case, there would be 150–166 years gained/lost from annual screening and 199–221 years gained/lost from biennial screening between age 40–49 (Table 2).

Net Benefit From Annual Versus Biennial Screening

Benefit/risk ratios for biennial screening are approximately 1.3 times higher than those for annual screening of women ages 40–49 because radiation risks from annual screening are twice that of biennial screening, whereas mortality reduction is only 1.5 times (36/24) higher. Of course, this observation does not necessarily imply that biennial screening is preferable. Net benefit, expressed as differences between lives saved and lives lost or as differences between years of life expectancy gained and years of life lost through screening, may be useful for comparing different screening regimens. Values for net benefit from annual screening shown in Table 3 are always approximately 1.5 times higher than the corresponding values for net benefit from biennial screening shown in Table 4.

Although subsequent annual or biennial screening after age 50 appears to have a substantial effect on benefit/risk ratios for screening women age 40–49 (Tables 1 and 2), such subsequent screening has relatively little effect on net benefit from screening women in their forties (Tables 3 and 4).

Radiation Risk and Other Risk Factors

Risk factors associated with radiation are incompletely known and, for some risk factors, may be extremely difficult to evaluate. For example, older age is a major risk factor for breast cancer, yet there is an inverse relationship between radiation sensitivity and age at exposure (*11*). Environmental factors are also hard to assess. For instance, although American women have a higher breast cancer incidence than Japanese women, probably due to diet and other environmental factors, absolute breast cancer risk from radiation is similar when both populations are compared, but relative risk factors are markedly different (*25*).

There are also possible genetic risk factors. One report claimed a fivefold or sixfold excess risk of breast cancer among blood relatives of patients with ataxia-telangiectasia who had received single or multiple diagnostic x-rays with an extremely low estimated dose to the breast glandular tissue of 1–9 mGy (*26*). A number of experts have expressed skepticism about these results, however, due to small sample size, inadequate assessment of radiation exposure, inconsistencies in results, presence of other confounding differences between the study and control groups, and incompatibility of this study with much larger studies showing no increase in breast cancer among women exposed to radiation after age 40 (*27–30*). Moreover, women who are heterozygous for the ataxia-telangiectasia gene represent less than 1% of the U.S. female population (*31*).

Inherited mutations in the BRCA 1 and BRCA 2 genes may be involved in 14% of breast cancers among women ages 40-49 and progressively lower percentages of breast cancers among older women (*31*). Meaningful studies of radiation sensitivity in women with inherited BRCA 1 and BRCA 2 mutations have not yet been performed and might not be feasible due to their very high baseline breast cancer incidence and the fact that they represent a relatively small proportion of the general population. Other factors, such as patient confidentiality and continued medical insurability, might also affect the ability to identify women with inherited gene mutations for these studies.

Table 3. Net lives saved due to annual mammographic screening of 100,000 women beginning at age 40 until age 49*

	Subsequent screening after age 50		
	None	Biennial	Annual
Lives saved due to screening	292	292	292
Lives lost due to radiation-induced breast cancers	8	4.9	4.4
Net lives saved	284	287.1	287.6

*Calculated using data and assumptions of Table 1.

Table 4. Net lives saved due to biennial mammographic screening of 100,000 women beginning at age 40 until age 49*

	Subsequent screening after age 50		
	None	Biennial	Annual
Lives saved due to screening	194	194	194
Lives lost due to radiation	4	2.4	2.2
Net lives saved	190	191.6	191.8

*Calculated using data and assumptions of Table 1.

Conclusion

For the general population of women ages 40–49, the theoretical radiation risk from screening mammography is extremely small compared with the established benefit from this life-saving procedure. Subgroup analysis of radiation sensitivity in high-risk women should not become a distraction from this overriding conclusion.

References

(*1*) Tokunaga M, Land CE, Tokuoka S, Nishimori I, Soda M, Akiba S. Incidence of female breast cancer among atomic bomb survivors, 1950–1985. Radiation Research 1994;138:209–23.

(*2*) Hrubec Z, Boice JD, Monson RR, Rosenstein R. Breast cancer after multiple chest fluoroscopies: second follow-up of Massachusetts women with tuberculosis. Cancer Res 1989;49:229–34.

(*3*) Miller AB, Howe GR, Sherman GJ, Lindsay JP, Yaffe MJ, Dinner PJ, et al. Mortality from breast cancer after irradiation during fluoroscopic examinations in patients being treated for tuberculosis. N Engl J Med 1989;321:1285–89.

(4) Shore RE, Hildreth N, Woodard ED, Dvoretsky P, Hempelmann L, Pasternack B. Breast cancer among women given x-ray therapy for acute postpartum mastitis. J Natl Cancer Inst 1986;77:689–96.
(5) Mattson A, Bengt-Inge R, Hall P, Wilking N, Rutqvist LE. Radiation-induced breast cancer: Long-term follow-up of radiation therapy for benign breast disease. J Natl Cancer Inst 1993;85:1679–85.
(6) Hancock SL, Tucker MA, Hoppe RT. Breast cancer after treatment of Hodgkin's disease. J Natl Cancer Inst 1993;85:25–31.
(7) Upton AC, Beebe GW, Brown JM, Quimby EH, Shellabarger C. Report of the NCI Ad Hoc Working Group on risks associated with mammography in the mass screening for the detection of breast cancer. J Natl Cancer Inst 1977;59:481–93.
(8) BEIR III Committee on the Biological Effects of Ionizing Radiation. The effects on populations of exposure to low levels of ionizing radiation. Washington (DC): National Academy of Sciences, 1980.
(9) National Institutes of Health Ad Hoc Working Group to Develop Radioepidemiological Tables. Report of the National Institutes of Health Ad Hoc Working Group to Develop Radioepidemiological Tables. NIH Publication No. 85-2748. Bethesda (MD): National Institutes of Health, National Cancer Institute, 1985.
(10) BEIR V Committee on the Biological Effects of Ionizing Radiation. Health effects of exposure to low levels of ionizing radiation. Washington (DC): National Academy Press, 1990.
(11) United Nations Scientific Committee on the Effeets of Atomic Radiation. Sources and Effects of Ionizing Radiation, UNSCEAR 1994 Report to the General Assembly with Scientific Annexes. New York: United Nations, 1994.
(12) Dvoretsky PM, Woodard E, Bonfiglio TA, Hempelmann LH, Morse IP. The pathology of breast cancer in women irradiated for acute postpartum mastitis. Cancer 1980;46:2257–62.
(13) Tokuoka S, Asano M, Tsutomu Y, Tokunaga M, Sakamoto G, Hartmann WH, et al. Histologic review of breast cancer cases in survivors of atomic bombs in Hiroshima and Nagasaki, Japan. Cancer 1984;54:849–54.
(14) Land CE. Estimating cancer risk from low doses of ionizing radiation. Science 1980;290:1197–1203.
(15) Feig SA, Ehrlich SM. Estimation of radiation risk from screening mammography: Recent trends and comparison with expected benefits. Radiology 1990;174:638–47.
(16) Feig SA, Dodd GD, Hendrick RE. Mammography risks and benefits. In: Radiation Protection in Medicine, Proceedings of the Twenty-eighth Annual Meeting of the National Council on Radiation Protection and Measurements, Proceedings No. 14. Bethesda (MD): National Council on Radiation Protection and Measurements, 1993:240–53.
(17) Law J. Risk and benefit associated with radiation dose in breast screening programmes—an update. Br J Radiol 1995;68:870–6.
(18) Mettler FA, Upton AC, Kelsey CA, Ashby RN, Rosenberg RD, Linver MN. Benefits versus risks from mammography: a critical reassessment. Cancer 1996;77:903–9.
(19) Smart CR, Hendrick RE, Rutledge JH III, Smith RA. Benefit of mammography screening on women ages 40 to 49 years. Current evidence from randomized controlled trials [published erratum appears in Cancer 1995; 75:2788]. Cancer 1995;75:1619–26.
(20) Committee and Collaborators, Falun meeting. Report of the meeting on mammographic screening for breast cancer in women aged 40–49, Falun, Sweden, March 1996. Int J Cancer 1996;68:693–9.
(21) Feig SA. Estimation of currently attainable benefit from mammography screening of women aged 40–49 years. Cancer 1995;75:2412–9.
(22) Geise RA, Palchevsky A. Composition of mammographic phantom materials. Radiology 1996;198:347–50.
(23) Young KC, Wallis MG, Ramsdale ML. Mammographic film density and detection of small breast cancers. Clin Radiol 1994;49:461–5.
(24) Gloeckler-Ries LA, Miller BA, Hankey BF, Kosary CL, Harras A, Edwards BK, editors. SEER Cancer Statistics Review, 1973–1991: Tables and Graphs, National Cancer Institute, NIH Pub. No. 94-2789. Bethesda MD, 1994; Section IV: Breast: 116–35.
(25) Land CE. Studies of cancer and radiation dose among atomic bomb survivors: the example of breast cancer. JAMA 1995;274:402–7.
(26) Swift M, Morrell D, Massey RB, Chase CL. Incidence of cancer in 161 families affected by ataxia-telangiectasia. N Engl J Med 1991;325:1831–6.
(27) Boice JD Jr, Miller RW. Risk of breast cancer in ataxia-telangiectasia [letter; comment]. N Engl J Med 1992;326:1357–1358; discussion 1360–1.
(28) Wagner LK. Risk of breast cancer in ataxia-telangiectasia [letter; comment]. N Engl J Med 1992:326:1358; discussion 1360–1.
(29) Hall EJ, Geard CR, Brenner DJ. Risk of breast cancer in ataxia-telangiectasia [letter; comment]. N Engl J Med 1992;326:1358–9; discussion 1360–1.
(30) Land CE. Risk of breast cancer in ataxia-telangiectasia [letter; comment]. N Engl J Med 1992;326:1359–61.
(31) Claus EB, Schildkraut JM, Thompson WD, Risch NJ. The genetic attributable risk of breast and ovarian cancer. Cancer 1996;77:2318–24.

Mammography Versus Clinical Examination of the Breasts

*Cornelia J. Baines, Anthony B. Miller**

Using published data from screening trials, this article compares two-modality (mammography and clinical examination) and single-modality (clinical examination alone) screening by evaluating cancer detection rates, program sensitivities, mode of cancer detection in two-modality screening, nodal status at time of detection, survival 10 years post-diagnosis, and breast cancer mortality 10 years after entry. Consistently, two-modality screening achieved higher cancer detection rates and program sensitivity estimates than either modality alone; mammography alone achieved higher rates than clinical examination alone; interval cancer detection rates between screening examinations were higher following clinical examination alone than mammography alone; single-modality screening with mammography failed to detect breast cancers identified by clinical examination alone; the sensitivity of mammography was lower in younger than older women, while the reverse was true for clinical examination; and mammography identified a higher proportion of node-negative breast cancer than clinical examination. We conclude that combining clinical breast examination with mammography is desirable for women age 40–49 because mammography is less sensitive in younger than older women. Careful training and monitoring are, however, as essential with clinical examiners as with mammographers. [Monogr Natl Cancer Inst 1997:22:125–129]

In countries where breast cancer is not a major priority and where funding for and expertise in screening mammography are scarce, clinical examination of the breasts as a single screening modality unquestionably deserves consideration. However, in North America, where breast cancer is a priority and mammography is relatively accessible, the real issue is not "screening mammography versus clinical examination," but rather "screening mammography with clinical examination versus screening mammography without clinical examination." Unfortunately, this issue is rarely addressed, probably due to two major factors: pervasive confidence in technology as a solution for most problems facing society; and the population's generally inflated view of the risks of getting breast cancer, of dying from breast cancer and, in particular, of benefiting from mammographic screening (*1,2*). Furthermore, when clinical breast examination is considered for inclusion in a mammography screening program, it is often dismissed for economic reasons. In general, mammography gets much attention and clinical breast examination is usually given short shrift, just as chest x-rays and electrocardiograms have diminished reliance on percussion and auscultation of the chest.

Several years ago, at a meeting on breast cancer, a speaker commented that "in an era when modern mammography is available, the use of clinical breast examination in screening is unethical and irrational." He raised an important issue. What is the role of clinical breast examination in screening for breast cancer? Is it dispensable? Answering these questions is difficult because there are few opportunities allowing valid comparisons of two-modality screening (mammography and clinical breast examination) with single-modality screening (clinical breast examination alone). The question may be particularly important for women age 40 to 49, for whom there is widespread controversy about the efficacy of mammography screening.

Data Sources

Of eight randomized controlled trials (RCTs) of breast cancer screening reported to date (*3*), the four Swedish RCTs used only mammography, leaving four that incorporated clinical breast examination in their protocol, namely the New York Health Insurance Plan (HIP) Study (*4*), the Edinburgh RCT (*5*), and the Canadian National Breast Screening Studies (CNBSS) I (*6*) and II (*7*) (Table 1). The manner in which these RCTs differed from each other must be understood. Respective ages at entry were 40–64 years, 45–64 years, 40–49 years, and 50–59 years. The intervention group in the HIP study and both CNBSS trials received annual two-view mammography and clinical breast examination. In the Edinburgh trial, the intervention group received two-view mammography with clinical breast examination at the first screen visit, one-view mammography with clinical breast examination at the third, fifth, and seventh screens, and clinical breast examination alone at the second, fourth, and sixth screens. Control groups in the HIP and Edinburgh trials received no screening at all. In CNBSS-I, the control group received a single clinical breast examination and thereafter depended on "usual care" in the community. They were followed annually by mailed questionnaire. In CNBSS-II, the control group received annual clinical breast examinations.

In addition to results from screening trials, clinical breast examination has been evaluated in case series (*8*) and in screening projects (*9,10*), all disadvantaged by the lack of an appropriate comparison group. In contrast to the case series, two screening projects—the Breast Cancer Detection Demonstration Project (BCDDP) (*9*) in the United States and the DOM Project in Utrecht (*10*), both of which used two-modality screening—do provide useful data on clinical breast examination for comparison purposes with the RCTs.

**Affiliation of authors:* University of Toronto, ON, Canada.

Correspondence to: Cornelia J. Baines, Associate Professor, Department of Public Health Sciences, University of Toronto, 12 Queen's Park Crescent West, 3rd Floor, Toronto, ON M5S 1A8, Canada.

Table 1. Available data sources for evaluating clinical breast examination*

Source (start date)	Design	Age at entry (year)	Study intervention	Frequency	Control intervention
HIP (1963) (*4*)	RCT	40–64	MA + CBE	q 1 y	No screening
Edinburgh (1979) (*5*)	RCT	45–64	MA + CBE (Rounds 1, 3, 5, 7)	q 2 y	No screening
			CBE (Rounds 2, 4, 6)	q 2 y	No screening
CNBSS-I (1980) (*6*)	RCT	40–49	MA + CBE	q 1 y	Single CBE
CNBSS-II (1980) (*7*)	RCT	50–59	MA + CBE	q 1 y	Annual CBE
BCDDP (1972) (*9*)	Project	37–74	MA + CBE	q 1 y	NA
Utrecht (1975) (*10*)	Project	50–64	MA + CBE	Variable	NA

*MA = mammography; CBE = clinical breast examination; q 1 y = annually; q 2 y = every two years.

The quality of the breast examination is also an important issue. Clearly, high performance standards are as important for clinical examination as for mammography. The CNBSS has established what competent clinical breast examination alone can achieve in terms of cancer detection (*11*). Recent research on the efficacy of breast self-examination (BSE) also reinforces the importance of high standards: benefit from BSE seems to be restricted to competent practitioners (*12,13*). We are not aware of any published document describing training, monitoring, or routine evaluation of clinical examiners in the HIP study, the BCDDP, or the Edinburgh trial. In the Utrecht Project, the clinical examination was performed by the radiological technologist at the time of mammography, and she used a cupped hand to palpate four quadrants of each breast (personal communication). This is in marked contrast to the method applied in the CNBSS, where the examiners were trained to visually examine the breasts, to palpate the whole breast (not just the cone), to use a systematic search pattern, and to apply the pads of their fingers. Women were examined both sitting up and lying down (*11*). Overall, the standards achieved by CNBSS clinical breast examination were high.

Approaches to Analysis

The constraints on evaluation of clinical breast examination arising from the various study designs are apparent in Table 1. Even so, the four RCTs and the two screening projects that combined clinical breast examination with mammography allow several approaches to evaluating the role of the former: cancer detection rates, program sensitivities, mode of cancer detection, nodal status at time of cancer detection, survival 10 years postdiagnosis, and breast cancer mortality 10 years after entry.

a) Cancer detection: Only the CNBSS allows detection rates to be compared for combined versus single-modality screening, since in the other two RCTs the comparison was screening versus no screening, and in the screening projects there were no control groups.

In CNBSS-I, for women 40–49 years on entry, two-modality screening can be compared with clinical examination alone, but only for cancers detected at the first screening visit and up to 12 months thereafter. (Because intervention women received their second screening examination 12 months after the first, subsequent comparisons on the role of clinical breast examination are not possible; one can only compare program outcomes with respect to breast cancer incidence and breast cancer mortality.) In contrast, in CNBSS-II, for women 50–59 years on entry, two-modality screening can be compared with clinical examination alone for four or five successive annual screening examinations.

Not only can breast cancer be detected as a direct consequence of a screening examination, it can also be detected in the interval between screening visits. Such "interval cancers" may only become detectable after the screening examination, or they may be missed by the screening process (in both cases, the screens are said to be "false negative"). Any evaluation of clinical breast examination must consider interval cancer rates.

b) Program sensitivities (detection method): Only the CNBSS studies yield sensitivity estimates for a screening protocol that includes clinical examination alone.

c) Mode of cancer detection: For women receiving two-modality screening, the proportions of breast cancer detected by mammography alone, by clinical examination alone, and by both simultaneously can be documented. This is possible within the intervention arm in the four RCTs and in the two screening projects.

d) Nodal status at time of detection: This offers yet another way to evaluate clinical examination. With the data available, comparisons of nodal status are possible according to both mode of detection within the intervention arms in all four RCTs and intervention versus control status in the two CNBSS trials. The latter is of greater relevance in evaluating clinical breast examination.

e) Survival postdiagnosis: For this approach, data from the three North American studies—CNBSS, HIP, BCDDP—relate survival to mode of detection for two-modality screening. Additionally, in the CNBSS, survival associated with single-modality screening can be reported.

f) Mortality from breast cancer 10 years after entry: CNBSS-I breast cancer mortality in cases with year-one screen and interval detections will be described because it offers the only opportunity to compare mortality following a single episode of two-modality screening to mortality following a single episode of single-modality nonmammographic screening in women age 40–49.

What Can Screening With Clinical Breast Examination Achieve?

a) Cancer detection: Unfortunately, there are only three trials in which clinical breast examination was conducted in the absence of mammography: the Edinburgh trial (*5*) and CNBSS I and II (*6,7*). It is clear from Table 2 that two-modality screening will detect more breast cancer than clinical examination

alone. Edinburgh detection rates for clinical examination at screening rounds 2, 4, and 6 were lower than the rates observed in the CNBSS for women age 50–59 (*7*). However, the Edinburgh women received mammography in rounds 1, 3, 5, and 7, and this may have depleted the breast cancers available for diagnosis by clinical examination in the following years. The rates might also be lower because the quality of the clinical examination did not match that in the CNBSS. A straightforward comparison of the Edinburgh and Canadian trials is clearly impossible.

The CNBSS-I detection rate for women age 40–49 screened with clinical examination alone at the first screening round was 2.46/1,000 (*6*), a rate exceeding the rates reported for the Swedish two-county and Stockholm mammography-alone trials, which were 2.09/1,000 and 2.06/1,000, respectively, in women age 40–49 at entry (*3*).

Interval cancer rates can be expected to be higher following screening with clinical breast examination alone compared to screening with mammography. At the first screen, for those age 40–49 at entry, CNBSS interval cancer rates were higher in women allocated to receive clinical breast examination only (1.11/1,000 women) than in those receiving two-modality screening (0.75/1,000 women). Given that CNBSS mammography achieves detection rates and sensitivity estimates that match other trials (*3*), it cannot be suggested that the comparison of CNBSS interval rates is unduly favorable to clinical examination. For women age 40–49, this raises the question, How important is the difference between interval rates in the two groups being compared?

b) Program sensitivity (detection method): Sensitivity estimates are higher for two-modality screening than for single-modality with mammography. For women age 40–49 on entry, the sensitivity of two-modality screening with two-view mammography was much higher in the CNBSS (81%) compared to single-modality with single-view mammography in the two-county study (62%) and in the first screen of the Stockholm study (53%) (*3*).

Only the CNBSS trials can compare the sensitivity of mammographic screening with screening by clinical breast examination alone (*11*). Comparing CNBSS sensitivity rates for two-modality screening to single modality with clinical breast examination in the two age groups, 40–49 at entry and 50–59 at entry, four observations can be made. First, two-modality screening achieved a higher sensitivity in older (88%) than younger (81%) women. Secondly, sensitivity estimates for clinical examination alone were slightly higher in younger women (68% vs. 63%). Thirdly, two-modality screening achieved a higher sensitivity in both older and younger women than single-modality screening with clinical breast examination. Fourth, in the CNBSS, the observed sensitivity for clinical examination alone in women age 40–49 is of the same order of magnitude (68%) as that observed in the two-county and Stockholm mammography-alone trials (62% and 53%, respectively) (*3*). However, the fact remains that mammography achieves higher detection rates than clinical examination alone (Table 3).

It may be advisable to use 1) two-modality screening for women age 40–49, based on lower mammography and higher clinical examination sensitivity estimates for this age group compared to older women and 2) single-modality, two-view mammographic screening for women 50 years and over, based on higher mammography sensitivity estimates and lower estimates for clinical examination for older women relative to younger women.

c) Mode of cancer detection: Table 3 displays the mode of cancer detection (for screen-detected tumors) in the intervention arms of the four RCTs and the two screening projects. The proportions detected by clinical breast examination alone vary from 3.3% in Edinburgh (*5*) to 44.7% in the HIP study (*15*). If one looks at the proportion of all clinically positive screening examinations, it varies from 44% for Utrecht to 74% for Edinburgh. The usefulness of clinical breast examination is demonstrated by the fact that there is no trial in which mammography identified all breast cancers. Although Edinburgh comes close at 96%, its provision of mammography every second year precludes results that can truly evaluate the role of clinical breast examination.

The data in Table 3 reinforce the advisability of adding clinical breast examination to mammography screening in younger women. In the CNBSS, for women age 40–49 allocated to mammography, clinical examination alone was the mode of cancer detection in 23.5%, compared to only 12% for women age 50–59 allocated to mammography.

d) Nodal status: Mammographically detected cancers are more likely to be node negative than those detected by clinical examination. Table 4 reveals not only that single-modality screening in CNBSS-I detected fewer cancers (55) at the first screening round than two-modality (86), but also that the latter is associated with a higher proportion of node-negative invasive tumors and a marginally higher proportion of node-positive tu-

Table 2. Screen detection rates/1000: two- versus single-modality screening*

Trial age intervention rounds	CNBSS				Edinburgh (*5*)	
	40–49 (*6*)		50–59 (*7*)		45–64	
	MA + CBE	CBE	MA + CBE	CBE	MA + CBE	CBE
1	3.89	2.46	7.20	3.45	6.15	—
2	1.74	—	3.74	1.95	—	1.75
3	1.99	—	2.48	1.28	3.15	—
4	2.38	—	3.14	0.89	—	0.85
5	1.84	—	2.84	1.64	3.33	—
6	—	—	—	—	—	1.03
7	—	—	—	—	3.08	—

*MA = mammography; CBE = clinical breast examination.

Table 3. Mode of cancer detection with two-modality screening*

Study	Age (year)	n	Percent detected at screening		
			MA only	CBE only	MA + CBE
HIP (1988) (*14*)	40–64	132	33.3	44.7	22.0
BCDDP (1987)† (*9*)	37–74	3548	35.5	7.9	53.3
Edinburgh (1990)‡ (*5*)	45–64	88	22.7	3.4	73.9
CNBSS† (*6*)	40–49	255	40.4	23.5	36.1
CNBS† (*7*)	50–59	325	53.2	12.0	34.8
Utrecht (1984)‡ (*10*)	50–69	196	55.6	9.7	34.6

*MA = mammography; CBE = clinical breast examination.
†All cancers.
‡Invasive cancers.

Table 4. Screen-1 nodal status of invasive cancers in patients age 40–49 from CNBSS*

Allocation nodal status	Annual MA + CBE		Single CBE	
	n	(%)	n	(%)
Node-negative	52	(60)	30	(54)
Node-positive	33	(38)	20	(36)
Status unknown	1	(2)	5	(10)
Total	86	(100)	55	(100)

*MA = mammography; CBE = clinical breast examination.

mors. Since equal numbers of women were assigned to the intervention and control arms, 25,214 and 25,216, respectively, it is appropriate to show frequencies rather than rates.

e) Survival postdiagnosis: Table 5 compares survival at 10 years postdiagnosis according to mode of detection for women with screen-detected breast cancer. Lead-time bias is not an issue here because what is being described are the survival rates associated with the three modes of detection possible in two-modality screening. The comparisons across these North American screening studies are impeded by unmatched age groupings for the cohorts, with younger women having a lower risk of breast cancer than older. CNBSS-I has the youngest cohort (age 40–49) (*6*) compared with the HIP Study (age 40–64) (*14*) and the BCDDP (age 34–74) (*9*). The highest survival rates are observed in the CNBSS for every mode of detection in women who received combined mammography and clinical examination. Because the CNBSS is the most recently conducted study of the three displayed, one contributing element may be better mammographic technology in the CNBSS compared to the two older trials. CNBSS 10-year survival postdiagnosis for women assigned to receive a single clinical examination matched that observed for women in the mammography arm who were detected by clinical examination only. Although the survival rates associated with detection by mammography alone in all three studies exceed those for the other two modes of detection, this is insufficient to prove that mortality from breast cancer has been reduced.

The major conclusion of the Boston case series (*8*), implausibly endorsed in the journal *Science* (*15*), was that five-year survival postdiagnosis was excellent at 95% for women whose breast cancers were detected by mammography alone, while that for women whose breast cancer was physically palpable was much lower at 74%. The usefulness of such case series, however, is limited by the lack of an appropriate comparison group, lead-time bias, and selection bias. Indeed, in CNBSS-1, seven-year survival for breast cancer patients age 40–49 detected by mammography alone was 95%, while that for women who did not receive mammography was 91% (*6*).

f) Deaths 10 years after entry: Much concern has been expressed about the asymmetric distribution of advanced breast cancer in CNBSS-I at the first screening round in women age 40–49 on entry (*16,17*), namely an excess of advanced breast cancer detected in the two-modality arm of the trial compared to the control arm. Table 6 displays the distribution of deaths that have occurred approximately 10 years after entry, in CNBSS-I women who had breast cancer detected either at the first screening round or in the first 12 months thereafter. For two-modality and single-modality groups, the distribution of breast cancer deaths in cases detected in the first year is now 21 versus 19, respectively, compared to 16 versus 10 at the seven-year follow-up (*6*). Including deaths in women with breast cancer due to other causes [all causes of death in breast cancer patients are verified by external panel review (*18*)], the totals are 22 and 21 for the two groups, respectively. This near equalization in distribution should lessen the persuasiveness of criti-

Table 5. Survival at 10 years by mode of detection—screen cancers only (%)*

Mode of detection	HIP (40–64) (*4*)	BCDDP (34–74) (*9*)	CNBSS (40–49) (*6*)	
	MA + CBE	MA + CBE	MA + CBE	Single CBE
MA only	77	85	93	NA
CBE only	59	76	84	86
MA + CBE	55	77	81	NA

*MA = mammography; CBE = clinical breast examination.

Table 6. Deaths due to invasive breast cancer 10 years* after entry among CNBSS subjects aged 40–49†

Allocation	MA + CBE	CBE only
Screen-1	15	9
Interval-1	6	10
Deaths due to other causes	1	2
Total	22	21

*One CBE screen death occurred at 10 y 3 m and one at 10 y 16 d.
†MA = mammography; CBE = clinical breast examination.

cism directed at the CNBSS (*16,17*), especially in light of similar patterns of mortality observed in other trials (*18*). The CNBSS mortality results are compatible with conclusions reached from meta-analyses, namely that benefit from screening women age 40–49 is slow to appear. The recently published external review of CNBSS randomization (*19*) by forensic experts found no evidence of subversion in CNBSS randomization procedures. An accompanying editorial (*20*) includes factual inaccuracies that have recently been corrected (*21*). This is not the first time that factual inaccuracies have been published (*18*).

Conclusions

Proponents of mammography screening in women age 40–49 have rightly said it is inappropriate to recommend clinical breast examination for screening in the absence of evidence. Certainly evidence from an RCT in North America comparing screening with clinical breast examination to no screening will never be available. Therefore, evidence on clinical breast examination from existing trials and projects must be examined. In fact, only the CNBSS allows comparative evaluation of clinical breast examination, and the comparison is with two-modality screening, not "no screening."

Because proponents of mammography have repeatedly called the CNBSS mammography "flawed" (*16,17*), the question arises, Are the achievements of clinical breast examination in the CNBSS enhanced because of "flawed mammography"? As has been reported before (*18*), there is much evidence to answer "no." The CNBSS has achieved results equal to or better than other RCTs with respect to successful randomization (which cannot be similarly documented for any other trial), cancer detection rates, prevalence/incidence ratio, and survival. In short, there is no persuasive evidence that "flawed" mammography enhanced the achievements of clinical breast examination observed in the CNBSS.

Unquestionably, by any of the parameters examined, screening with mammography, alone or in combination with clinical exam, performs better than clinical breast examination alone. The differences described may be smaller and possibly less important than many would predict, with one major exception: cancer detection rates. These are always considerably higher when mammography is used. Nevertheless, this may not be an unqualified benefit given the likelihood of overdiagnosis (*22,23*).

Because two-modality screening out-performs mammography alone, there is a role for clinical breast examination in breast screening if women are to gain the most benefit from screening. It has long been known that biopsy of a palpable mass should not be deferred because of negative mammograms (*24*). With mammography alone, lumps will be overlooked, especially in younger women.

As with mammography, breast examination technique must be excellent in order to be useful. And excellence can be achieved. It has been demonstrated that medical school curricula could be revised to enhance clinical breast examination competence among medical students (*25*) and that educational programs can effectively improve examination competence among health professionals (*26*). The need to achieve excellence should not be a deterrent to clinical breast examination any more than it has been to mammography. If clinical breast examination is to be employed in screening, examiners will need to be carefully trained and monitored. If the costs of a screening program must be limited, one could recommend that clinical breast examination should, at the very least, be part of the screening protocol for women under age 50 because, at that age, the sensitivity of mammography is lower than in later years.

References

(*1*) Black WC, Nease RF Jr, Tosteson AN. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. J Natl Cancer Inst 1995;87:720–31.

(*2*) Baines CJ. Women and breast cancer: is it really possible for the public to be well informed? [editorial]. Can Med Assoc J 1992;146:2147–8.

(*3*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.

(*4*) Shapiro S, Venet W, Strax P, Venet L, Roescer R. Ten- to fourteen-year effect of screening on breast cancer mortality. J Natl Cancer Inst 1982;69:349–55.

(*5*) Roberts MM, Alexander FE, Anderson I, Chetly U, Donnan PT, Forrest P, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241–6.

(*6*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years [published erratum appears in Can Med Assoc J 1993;148:718]. Can Med Assoc J 1992;147:1459–76.

(*7*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: II. Breast cancer detection and death rates among women aged 50–59 years [published erratum appears in Can Med Assoc J 1993;148: 718]. Can Med Assoc J 1992;147:1477–88.

(*8*) Stacey-Clear A, McCarthy KA, Hall DA, Pile-Spellman E, White G, Hulka C, et al. Breast cancer survival among women under age 50: is mammography detrimental? Lancet 1992;340:991–4.

(*9*) Seidman H, Gelb SK, Silverberg E, LaVerda N, Lubera JA. Survival experience in the Breast Cancer Detection Demonstration Project. CA 1987;37: 258–90.

(*10*) DeWaard EF, Collette HJA, Rombach JJ, Banders-van Halewijn EA, Honing C. The DOM project for the early detection of breast cancer, Utrecht, The Netherlands. J Chron Dis 1984;37:1–44.

(*11*) Baines CJ, Miller AB, Bassett AA. Physical examination. Its role as a single screening modality in the Canadian National Breast Screening Study. Cancer 1989;63:1816–22.

(*12*) Newcomb PA, Weiss NS, Storer BE, Scholes D, Young BE, Voigt LF. Breast self examination in relation to the occurrence of advanced breast cancer. J Natl Cancer Inst 1991;83:2605.

(*13*) Harvey BJ, Miller AB, Baines CJ, Corey PN. A nested case-control study of breast self examination practice. Can Med Assoc J. In press.

(*14*) Shapiro S, Venet W, Strax P, Venet L. Current results of the breast cancer screening randomized trial: the Health Insurance Plan (HIP) of Greater New York Study. In: Day NE, Miller AB, editors. Screening for Breast Cancer. Toronto: Hans Huber, 1988:3–15.

(*15*) Good news on mammograms. Science 1992;258:739.

(*16*) Tarone E. The excess of patients with advanced breast cancer in young women screened with mammography in the Canadian National Breast Screening Study. Cancer 1995;75:997–1003.

(*17*) Kopans DB, Feig SA. The Canadian National Breast Screening Study: a critical review. AJR Am J Roentgenol 1993;161:755–60.

(*18*) Baines CJ. The Canadian National Breast Screening Study: a perspective on criticisms. Ann Int Med 1994;120:326–34.

(*19*) Bailar JC III, MacMahon B. Randomization in the CNBSS: a review for evidence of subversion. Can Med Assoc J 1997;156:193–9.

(*20*) Boyd NF. The review of randomization in the Canadian National Breast Screening Study. Is the debate over? Can Med Assoc J 1997;156:207–9.

(*21*) Cohen M, Kaufert P, MacWilliam L, Tate R. Checking random assignment with claims data. Can Med Assoc J 1997;156:1269–70.

(*22*) Fletcher SW. Breast cancer screening among women in their forties: an overview of the issues. Monogr Natl Cancer Inst 1997;22:5–9.

(*23*) Andersson I, Janzon L. Reduced breast cancer mortality in women under age 50: Updated results from the Malmö Mammographic Screening Program. Monogr Natl Cancer Inst 1997;22:63–7.

(*24*) Winchester DP. Physical examination of the breast. Cancer 1992;69: 1947–9.

(*25*) Campbell HS, McBean M, Mandin H, Bryant H. Teaching medical students how to perform a clinical breast examination. Acad Med 1994;69:993–5.

(*26*) Campbell HS, Pilgrin CA, Fletcher SE, Morgan TM, Lin S. Improving physicians' and nurses' clinical breast examination. A randomized controlled trial. Am J Prev Med 1991;7:1–8.

The Psychosocial Consequences of Mammography

*Barbara K. Rimer, Leslie G. Bluman**

Increasing numbers of mammograms being performed in the United States will be accompanied inevitably by an increasing number of false positives. According to reliable estimates from a survey of radiology facilities, U.S. women in their forties experience close to one million false positive mammograms every year. To determine the impact of false positive mammograms and the broader psychological impact of mammography, we conducted literature searches of Medline, CancerLit, and PsycInfo. We identified nine studies examining the impact of false positive mammograms. Most found short-term increases in such psychological measures as anxiety, distress, and intrusive thoughts. One study found substantial effects on these measures three months after an abnormal mammogram. Another study found an 18-month impact on anxiety. Few studies have used behavioral outcomes, but one reported overpractice of breast self-exam among women who had received false positive results. Another found no reduction in adherence to mammography among women who have had an abnormal test. The more general mammography literature suggests that many women are anxious about mammography before the exam; women with lower levels of education, African Americans, and women with a family history of breast cancer may be more vulnerable to distress. Unfortunately, this literature suffers major limitations, such as small sample sizes, inconsistent and sometimes inappropriate measures, variations in the time frames for measurement, few studies with women aged 40–49, and a paucity of U.S. research. More research is needed to characterize at-risk women and to test interventions designed to reduce the negative impact of abnormal mammograms. Improved communication is also needed throughout the entire mammography process. [Monogr Natl Cancer Inst 1997:22:131–138]

Mammography use has increased dramatically in the past 10 years. In 1987, the National Health Interview Survey (NHIS) found that only about one-third of U.S. women had ever had a mammogram, and only 17% had had one in the preceding year (*1*). By the 1992 NHIS, 70% of women aged 40–49 reported having had a mammogram, 36% of U.S. women reported having been screened recently, and 35% said they had had one in the last year (*2*). These increases carry an inevitable burden of false positives and false negatives. The number of false positive mammograms received by U.S. women may be as high as 2.75 million, based on the 11% false positive rate found in a survey of community facilities (*3*) and an estimate of about 25 million screening exams performed annually in the United States (Fletcher S, personal communication). If about 35% of the total mammograms performed each year, based on NHIS data (*2*), are in women aged 40–49, and 11% are false positives (*3*), 960,000 abnormal mammograms could occur annually in this age group. Thus, it is appropriate to consider both the negative and positive consequences of receiving an abnormal result. These consequences could be factored into the overall mammography benefit-risk ratio for women of different ages.

Studies have examined outcomes of the general mammography experience, such as anxiety, distress, depression, excessive fear of cancer, subsequent practice of breast self-exam (BSE), and adherence to recommended mammography schedules or other follow-up procedures (*4–6*). Some studies have focused more specifically on the psychological sequelae of abnormal mammograms. One concern is that the experience of an abnormal mammogram may not only cause psychological reactions, such as severe anxiety and distress, but also could act as a negative reinforcer deterring women from subsequent mammograms. Yet for a field as large as breast cancer screening, there has been surprisingly little study of the psychological consequences of mammography, especially compared to the amount of research on the psychosocial barriers to mammography. Moreover, most of the studies have been conducted in Europe, leading to an uncertain ability to generalize results to the United States.

This review focuses on the psychosocial consequences of abnormal mammograms. Some consideration of the more general mammography experience is presented in order to place reactions to abnormal mammograms in context. Where published, reports about interventions to help women cope with the abnormal mammography experience also are included. The larger issue of compliance with recommended follow-up for abnormal mammograms, while important, is beyond the scope of this report.

The literature on the psychosocial consequences of abnormal exams is extremely limited. Three separate searches of Medline, CancerLit, and PsycInfo between October 1996 and December 1996 identified fewer than 30 discrete articles, some of which were anecdotal reports or tangential to the topic. We also wrote to investigators who are conducting research in this area to identify in-press articles—none were forthcoming. This review fo-

**Affiliation of authors:* Duke University Medical Center, Comprehensive Cancer Center, Cancer Prevention, Detection and Control Research Program, Durham, NC.

Correspondence to: Barbara K. Rimer, Dr.P.H., Duke University Medical Center, Cancer Prevention, Detection and Control Research Program, Trent Drive, Hanes House, Box 2949/DUMC, Durham, NC 27710.

See "Notes" following "References."

cuses on published reports about responses to abnormal mammograms and about the psychosocial consequences of mammograms. We included only articles that provided data and were not exclusively case reports, single group analyses, or exploratory studies. Because of the relatively undeveloped nature of the field, nonexperimental studies and cross-sectional surveys were included. However, anecdotal and case reports were excluded.

Psychological Impact of an Abnormal Mammogram

As Paskett and Rimer (*6*) have discussed, there can be several consequences of abnormal medical tests. These include labeling, psychologic distress, and noncompliance with evaluation or treatment recommendations. Most of the published research has focused on psychological reactions, such as intrusive thoughts, worry, and distress.

Nine published reports (summarized in Table 1) have examined various aspects of the abnormal mammogram experience (*4,7–14*). There have been other reports, but many of these reports are largely exploratory, and the results are limited by small samples, often collected in a nonrandom manner. Most of the studies on which we focus have included a range of ages; therefore, the results cannot be examined separately for women aged 40–49. The outcomes have included various measures of distress, anxiety, hostility, effect on BSE practice, and impact on adherence to mammography. One study also obtained endocrine and immunologic measures. This first group of studies includes only those empirical reports in which at least some subset of the sample received abnormal results. Women with breast cancer were excluded from all studies except the report by Ellman, Angeli, Christians, et al. (*4*).

Two studies (*7,12*) found short-term negative emotional reactions in women who have had abnormal thermograms or mammograms, but the sample sizes were quite small. Bull and Campbell (*8*) sent questionnaires to 750 women prior to breast cancer screening and subsequently to women with normal findings and those who required follow-up procedures as a result of abnormal exams. There was no increase in general levels of depression or anxiety in any of the groups; however, there was a significant increase in the overpractice of BSE among women who required special assessments, especially biopsies, as a consequence of abnormal exams. This is of concern, since overpractice of BSE may diminish the ability to detect subtle changes in breast tissue (*15*).

Ellman et al. (*4*) compared different subgroups of women in a sample of 733 women in the UK Trial of Early Detection of Breast Cancer. Three months after attendance at a recall clinic, the same proportion (19%) of women with false positive results and with routine screening experienced anxiety. Women with symptomatic benign conditions had anxiety scores that were elevated three months later. Although there was a short-term increase in anxiety among the false positive group, it was not sustained. Sutton, Saidi, Bickler, et al. (*13*) analyzed data from the National Health Service Breast Screening Programme on 306 attenders and 100 nonattenders; however, only 24 women were in the false positive category. Anxiety was highest at baseline, but, in general, the women were not overly anxious. On retrospective analysis, women with false positives recalled feeling more anxious than negative screenees. Another examination of women in this program found significant increases in worry, and physical, emotional, and social dysfunction was found among the women who were recalled. Distress was higher in women with a personal history of breast problems or a family history of breast cancer (*14*).

The false positive experience may affect women's perceptions about mammography and, thus, make them anxious about future exams. In one of the larger studies (nearly 300 women), Gram, Lund, and Slenker (*9*) and Gram and Slenker (*10*) found that women in the Tromso, Norway, screening program who had false positives were retrospectively more likely than negative screenees to rate mammography as unpleasant or both painful and unpleasant. Moreover, they found the effects on increased anxiety to be long lasting: 18 months after screening, 29% of women with false positives reported anxiety compared to 13% of those with negative results. A small proportion (5%) of the women with false positives described the experience as the worst in their lives. About 11% said that their capacity for work was affected during the waiting period, but 44% said that the abnormal mammography experience had an overall positive impact on their lives.

Lerman and colleagues (*11,16*) evaluated women's psychological responses to abnormal mammograms and the effect on mammography adherence. The authors assessed psychological responses and subsequent adherence to mammography among 300 women in an Independent Practice Association-model health maintenance organization (HMO) who had had mammograms with varying levels of suspicion. The degree of mammogram suspicion was significantly related to the strength of the adverse outcome. Women with more suspicious abnormal mammograms reported significantly elevated levels of distress, and their mammography-related anxiety and breast cancer worries interfered with their moods and functioning: in the high-suspicion group, 47% had mammography-related anxiety and 63% had worries about breast cancer; such worries affected the moods (38%) and daily functioning (27%) of these women. Women with high and low levels of impairment were less likely to practice BSE than those with moderate impairment. Intentions to get mammograms in the next year increased directly with the level of mammogram suspicion. Most women with abnormal mammograms obtained their next mammogram on schedule. These data suggest that even when the results of an abnormal mammogram are shown not to be cancer, some women experience negative sequelae. However, this study was conducted among women aged 50–74, and it is not clear to what extent the results would be similar among women aged 40–49. Nevertheless, this is one of the largest and most well-controlled studies of the abnormal mammography experience to date.

Overall, the studies indicate that false positives have a moderate but reasonably consistent effect on such psychological measures as anxiety, worry, and distress. The majority of studies found statistically significant short-term increases in worry and/or distress. In several studies, about one-fifth or more of the women reported a negative effect of the abnormal mammogram on their daily functioning. Few studies have included longer-term impact measures. Lerman, Trock, Rimer, et al. (*11,16*)

found a substantial impact at three months after the abnormal mammogram result. Gram and Slenker (*10*) found significant anxiety 18 months after an abnormal result. In one study, the false positive event seemed to cause overpractice of BSE. Again, few studies have included behavioral outcomes. It is not possible to determine from these studies the impact of abnormal mammograms or the duration of negative sequelae.

Psychosocial Consequences of Mammography

A small body of literature includes studies of the psychological consequences of mammography in general. These studies are summarized in Table 2 (*17–20*). We did not consider the larger literature that is based primarily on retrospective accounts of the mammography experience.

Fine, Rimer, and Watts (*18*) interviewed 250 women immediately after they had mammograms: 60% of the women were anxious about having a mammogram, and 20% were extremely anxious. African-American women were significantly more anxious than white women, and those with a high-school education or less were significantly more anxious than those with more education. Some of this anxiety seemed to be due to a lack of information about what to expect. Baines, To, and Wall (*17*) assessed reactions to mammography among active respondents as part of the Canadian National Breast Screening Study (NBSS). Only 5.4% of the women said they were anxious about their mammograms, but the majority of those who responded this way said it was because of an abnormal referral. In a large sample (over 2,000 women), Walker, Cordiner, Gilbert, et al. (*20*) found that prior to screening, nearly 20% of the women had clinically significant anxiety scores and 6% had clinically significant depression scores. These scores decreased significantly between baseline and screening. Some women reported such adverse effects as difficulty sleeping, inability to concentrate, and inability to relax or feel happy during the week before screening.

One study (*19*) with a small sample (n = 53) indicated that women with a high familial risk of breast cancer had significantly higher levels of both acute and nonspecific distress and avoidant and intrusive thoughts after mammography when compared to normal-risk women. These results persisted one month after the normal report. The impact of family or other risk factors on response to mammography should be investigated further, since these women are likely to be advised to start mammography at a younger age.

Thus, the evidence from these studies suggests that a substantial proportion (20%–60%) of women are anxious about mammography before their exams; in some cases, the evidence was clinically significant. This baseline level of anxiety, then, could be exacerbated by abnormal results. Some women seem to be more adversely affected than others by the mammography experience. Among those more vulnerable to distress were African-American women, those with lower levels of education, and those with a family history of breast cancer. These may be the same women who will have more negative reactions to the abnormal mammography result, but more information is needed. Many women clearly would benefit from better preparation. Fine et al. (*18*), for example, found that anxiety was higher in women who felt less prepared for the mammogram.

Interventions to Reduce Negative Psychosocial Consequences and to Improve Coping

There has been scant research on interventions designed to reduce anxiety and distress and to improve coping after an abnormal result. In one of the few studies in this area, Lerman, Ross, Boyce, et al. (*21*) sent women in an experimental condition a booklet designed to improve adherence to the subsequent mammogram following an abnormal test. The brief psychoeducational booklet resulted in a statistically significant 13% increase in adherence to the subsequent mammogram.

Discussion

The research base on the psychosocial consequences of mammography, in general, and abnormal mammograms, in particular, is extremely limited. There are major methodological deficiencies among the published research studies. The investigators have studied different age groups and different time intervals and used a range of measures, measurements, and outcomes. In many cases, the sample sizes were so small as to render the results primarily exploratory. Some investigators have assessed responses immediately after the abnormal experience; others have used different time points. There is little consistency in the use of measures, and the selections rarely have been justified. Only two of the above-mentioned studies—those by Bull and Campbell (*8*) and by Walker et al. (*20*)—utilized a common measure to assess the psychosocial impact of mammography, and their samples were quite different. These studies incorporated the Hospital Anxiety and Depression Scale (HADS) to assess the effect of mammography on depression and anxiety. Overall, women attending for routine mammography experienced mean reductions in anxiety of 2.7% (*20*) and 10.9% (*8*) following the mammogram. Corresponding reductions in depression were 10.3% (*20*) and 15.4% (*8*).

The lack of a common set of measures across studies makes it inappropriate to conduct formal meta-analyses. Without a standardized measure, such as an effect size, it is difficult to compare the results of one study to another. Moreover, often the measures themselves are inappropriate. For example, general distress may not be as sensitive a measure as screening-related distress (*14*).

Different levels of support have been provided to help women cope with the abnormal experience, thus serving as a potential confounder. Women's reactions also may be affected by how the results are communicated. The generalizability of results may also be limited by the fact that most of the studies have been conducted in European countries where health care is provided free by the government, invitations are issued for mammography, and psychosocial support seems more likely to be provided.

Thus, it is difficult to reach clear conclusions about the impact of mammography or abnormal mammograms on such outcomes as anxiety, distress, or adherence to recommended breast screening. Among some women, there does seem to be short-term distress, and at least one study shows that the level of distress is related to the index of mammogram suspicion. The effects are

Table 1. Psychological responses to abnormal mammograms

Authors	Sample size	Age	Methods	Time of measurement	Results
Bartolucci G, Savron G, Fava GA, Grandi S, Trombini G, Orlandi C. 1989 (*7*)	50 patients who had a normal thermogram, 20 patients for whom there was an abnormal thermogram that turned out not to be cancer	Group 1: Mean = 38.2 Range = 17–61 Group 2: Mean = 48.8 Range = 41–61	Consecutive unselected women attending breast screening clinic in Italy. SAQs. RR not available.	Group 1: Immediately prior to thermography and then 3 to 4 days later, after learning of normal result Group 2: SAQ administered before mammogram, which followed abnormal thermography and 3 to 4 days later, after learning of normal results	Patients showed significant decreases in anxiety (p<.001), depression (p<.001), somatic arousal (p<.01), worry about illness (p<.05), concern about pain (p<.05), and fear of dying (p<.01) after hearing the normal results. There was a further decrease in anxiety and concern about pain when women learned of normal mammogram (p<.05). Thus, the authors noted that the experience entailed significant emotional arousal.
Bull AR, Campbell MJ. 1991 (*8*)	1125 women	All over age 50	Screening reactions were assessed at invitation, mammogram, attendance at special clinic for abnormal follow-up, and surgical biopsy in the UK. Women at the first stage were selected from six general practices in screening programs. Subsequent normal samples were drawn for the three next stages. SAQs. RR = 76%*	Invitation, mammogram, attendance at follow-up clinic, and biopsy	Significant increase in frequency of BSE occurred as index of abnormality increased (p<.001). After screening, 29/226 practiced BSE 1 or more times per week; 64 had increased BSE and 26 had decreased BSE. No significant differences in anxiety were found between the groups; 10% in abnormal groups said screening had left them more anxious; 10% of biopsy group had increased BSE to more than 1 time per week. Authors coneluded that the psychological effects were of note in women who needed biopsy.
Ellman R, Angeli N, Christians A, Moss S, Chaimberlain J, Maguire P. 1989 (*4*)	733 women	45–71	GHQ administered to 302 women attending routine screening, 300 women attending for follow-up of positive results, and 150 women with breast symptoms that were benign. Women were recruited on a weekly basis from clinics in the UK. RR = 94.7%* (women approached who completed both questionnaires)	At clinic before seeing doctor, women completed SAQ, then 3 months later, questionnaire was administered to women in their homes.	Women in the false positive and symptomatic benign abnormalities groups had significantly greater anxiety scores than those having routine screening (p<.02 and p<.002). 3 months later, the FP and routine group had the same level of anxiety, but this anxiety was significantly decreased among both groups (p<.005, p<.05).
Gram IT, Lund E, Slenker SE. 1990 (*9*)	126 women with false positives and 152 women with normal exams	40 and older	Women in the Tromso Screening Program, Tromso, Norway, were mailed SAQ six months after screening mammogram. Questionnaires were also sent to non-attenders and a community sample. In-person interviews conducted 18 months following screening. RRs = 79% (study group), 73% (comparison group)	After screening (same time for non-screened women) and 18 months later	29% of women with false positives reported anxiety 18 months after the event compared to 13% of those with negative results (p = .001). 5% described FP as the worst thing they had ever experienced. 18 months later, the majority of FPs reported the same quality of life as those with negative exams. Women's perceptions of the work-up period were longer than those documented in hospital files (p = .05). 63% said that they were anxious compared to 16% in the reference group. 11% in the recall group said that they had less capacity for work until they learned of their results. 44% said that the workup experience had an overall positive impact on their lives.

Table 1—Continued

Authors	Sample size	Age	Methods	Time of measurement	Results
Gram IT, Slenker SE. 1992 (*10*)	Negative screens (NS) = 209, False positives = 160, Non-attenders = 178, Population sample = 164	Median = 46 Range = 40–61	As part of the third Tromso, Norway, study, all abnormals who did not have cancer were identified, along with a sample of negative screenees and non-attenders and a random population sample. SAQs. RRs: 84% (screened negatives), 89% (false positives), 38% (non-attenders), 66% (population sample)	After mammography was completed; exact timing not available	Significantly more women in the false positive group than in the NS group reported the mammogram to be unpleasant (26%) or both painful and unpleasant (11%) (p<.01). Among the FP group, women who had been anxious about breast cancer at previous exams were more likely to be anxious at the current mammogram (p<.001).
Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. 1991 (*11*)	121 women with normal findings, 119 with low suspicion mammograms and 68 with high suspicion mammograms but not breast cancer (N = 308)	Mean = 58	Women were selected from an HMO pool of women in Pennsylvania and New Jersey who had recent mammograms and had not been diagnosed with breast cancer. Subjects were interviewed by phone. RR = 85%	3 months from mammogram	47% of women with high suspicion mammograms had mammogram-related anxiety, and 63% had worries about breast cancer. 38% of women said that their worries affected their mood, and 27% said that their daily functioning was affected. 41% of those with high suspicion findings compared to 28% of normals said they were at least somewhat worried about breast cancer. A decrease in concerns about breast cancer decreased chances of subsequent mammography adherence among all groups. Most women had subsequent mammograms on schedule.
Lidbrink E, Levi L, Pettersson I, Rosendahl I, Rutqvist LE, de la Torre B, et al. 1995 (*12*)	45 women who were recalled for 3-view mammographic exams and did not have breast cancer	NA	36 women were told of normal findings 1 hour after mammograms; the other nine were told one week later. The study, which took place in Sweden, includes not only psychological but also endocrine and immunological measures. SAQs. RR = 98% (volunteered), RR = 92%* (after women with breast cancer excluded)	2 different measurements; immediately after mammogram and three weeks after they were determined free of breast cancer. Long term (6 and 12 months) follow-up on 10 randomly chosen women	The mean mood score was lower at time 1 than time 2 (p<.05); no differences in endocrine or immunologic function were found. Emotion-focused copers had higher cortisol levels than problem-focused copers, suggesting greater stress. The authors speculated that the short waiting period may have attenuated the results.
Sutton S, Saidi G, Bickler G, Hunter J. 1995 (*13*)	Two overlapping samples. Sample A included 795 women who were due for screening at a mobile unit and returned questionnaires at 2 times. Sample B included 732 women who attended clinic during 3-month period and provided complete data. 306 attenders common to both samples and 100 non-attenders from Sample A were included. Only 24 FPs in all.	Mean = 58	This study had a prospective design with a retrospective analysis of anxiety and was conducted in the UK. SAQs. Sample A RR = 53%* (completed both questionnaires) RR = 27%* (included in analysis) Sample B RR = 84% (provided adequate data) RR = 35%* (included in analysis)	Questionnaires at three points: baseline, screening visit, and nine months later	Main analyses were on 306 attenders and 100 non-attenders. There was no significant difference in anxiety pre- and postscreening. Younger women were significantly more anxious (p<.01). On retrospective analysis, women with false positive results recalled feeling more anxiety at every stage as well as more pain and discomfort.

Table 1—Continued

Authors	Sample size	Age	Methods	Time of measurement	Results
Swanson V, McIntosh IB, Power KG, Dobson H. 1996 (*14*)	1285 women	50–64	SAQs were used to assess anxiety, concern about breast problems and other effects on women invited to the UK National Health Service Breast Screening Program. RRs = 49% (women invited for screening), 68% (women attending for mammography)	Baseline and after screening	56% of the women who attended screening reported reduced anxiety as a result of screening, while 13% reported increased anxiety. Women with a family history of breast cancer or breast disease tended to be more worried. Mammography did not increase anxiety among those not previously worried. There was a significant increase in worry and physical, emotional and social dysfunction in the group of women who were recalled ($p \leq .05$) and assessed at the time of recall.

RR = response rate, * = response rate calculated by reviewers; SAQ = self-administered questionnaire; GHQ = general health questionnaire.

relatively modest but not insignificant, with most studies indicating, not surprisingly, a significant increase in anxiety among women with abnormal results. While the majority of women do not suffer short-term harm, there seems to be a small group of women who are affected adversely. As Gram et al. (*9*) showed, the increased level of anxiety persisted 18 months after screening. Thus, the sequelae seem to be largely psychological—effects on such variables as worry, distress, and intrusive thoughts. To date, there is no evidence of a negative impact on subsequent mammography adherence, but only one study included this as a major outcome.

There is a need for rigorous research that includes sufficient numbers of women aged 40–49. Sample sizes should be adequate enough to conduct subgroup analyses by race and age. It would be useful to determine how long any negative effects persist after an abnormal mammogram. Research should be conducted in the United States, where cost may be a factor in response to the abnormal experience and where there is not a national health care system. Ideally, some studies would use telephone or in-person interviews to avoid the limitations of self-administered questionnaires (*22*). It would be helpful to obtain information about whether women missed time from work or usual activities in order to calculate the indirect costs and impacts of abnormal mammograms on women's lives.

It is critical to characterize the women who may be more likely to suffer adverse effects. As Swanson, McIntosh, Power, et al. (*14*) caution, it is important to recognize the diversity of responses when examining the impact of screening programs. Considering the effect on larger populations may mask substantial subgroup differences. If women at high risk for problems in coping can be identified, they can be provided with intervention in a proactive manner. There is some suggestion that women with a strong family history may be affected more negatively by an abnormal mammogram (*14,19*), but there are few data. Lerman, Daly, Sands, et al. (*23*) found an inverse association between psychological distress and family history among women with a family history of breast cancer. Lerman, Lustbader, Rimer, et al. (*24*) also found that high-risk women who were very anxious did not benefit from a risk-counseling program. So, at least among some women, there is reason for concern. Clearly, anxiety can interfere with learning. More investigation of this group is essential, since the current activity in genetic testing for cancer susceptibility is likely to result in more younger women having mammograms, with the inevitable consequence of more abnormal results.

It is not known to what extent the negative psychosocial sequelae of mammography might affect follow-up recommendations for additional tests or delay in seeking care for potential cancer symptoms (*5*). Noncompliance with follow-up recommendations continues to be a problem (*25*). Moreover, there is no information on the cumulative impact of more than one abnormal mammogram. There is reason to hypothesize that a second or third abnormal result could be especially distressing, but there are no data in this area.

The results of the Fine et al. (*18*) study suggest that better communication is needed throughout the process. This should begin with preparation for the mammogram. Moreover, anything that can be done to minimize the time between follow-up procedures and communication of results to women probably will reduce adverse effects (*4,9,26*).

The impact of brief psychoeducational interventions and other interventions designed to help women cope with the abnormal experience should be investigated. Researchers should test the efficacy of different strategies, not only to communicate abnormal findings, but to help women cope with the anxiety that occurs during the waiting period and thereafter. Only a subset of women are at risk for extreme anxiety, but there must be a mechanism by which to identify them and provide them with the needed support. Lerman et al. (*21*) demonstrated a 13% increase in mammography adherence with a minimal type of mailed psychoeducational intervention. This is extremely promising and suggests that low-cost, low-intensity interventions may have some value in faeilitating effective coping in response to the abnormal mammography experience. Telephone counseling has been used effectively in a number of health-related areas (*27*) to assist women who have particular difficulty in coping. It is not known whether women in their forties would have different intervention needs than women aged 50 and older.

Table 2. Psychological aspects of the mammography experience

Authors	Sample size	Age	Methods	Time of measurement	Results
Baines CJ, To T, Wall C. 1990 (*17*)	2299 women	40–59 at date of entry	SAQs used to assess attitudes after participation in the Canadian NBSS. After screening was completed, RR = 82%.	At completion of screening	After screening, only 5.4% said they were anxious; 15% said they were neither reassured nor anxious. Of women who reported anxiety, 60% said it was because of abnormal referral.
Fine MK, Rimer BK, Watts P. 1993 (*18*)	255 women	Mean = 52.8	Interviews with women in Philadelphia right after mammograms. An inception cohort was obtained through radiology centers. RR not available.	Immediately after mammogram	60% of the women said they felt anxious about their mammograms; one-third were quite a bit or extremely anxious. 71% of African-American women were anxious compared to 41% of white women (p<.0001). First time mammograms were more stressful than subsequent mammograms (p = .002). 83% of women with less than a high school education reported anxiety compared with 54% of women who had a high-school education or more (p = .001). 16% of all women were very worried about the result.
Valdimarsdottir HB, Bovbjerg DH, Kash KM, Holland JC, Osborne MP, Miller DG. 1995 (*19*)	58 women (none with abnormal reports were allowed)	Risk group Mean = 43.1 Compar. group Mean = 39.3	Women with a family history were recruited through a high risk clinic in New York, NY (n = 26). A comparison group was recruited from the community (n = 32). Measures were obtained at baseline and two time points after screening. SAQs. RR = 81%* (high-risk women), 84%* (comparison group, eligible and agreed to participate)	One month after screening	Acute distress was significantly higher prior to mammography as compared to 2 follow-ups (p = .005). The total mood disturbance decreased in the risk group but not in the comparison group (p<.006). Nonspecific psychological distress was higher in the risk group (p = .04). They also had higher levels of intrusive thoughts about breast cancer (p = .009) and avoidant thoughts (p = .006) even one month after normal result.
Walker LG, Cordiner CM, Gilbert FJ, Needham G, Deans HE, Affleck IR, et al. 1994 (*20*)	1635 women completed questionnaires at both baseline and screening	NA	Women eligible for the national screening program in Scotland completed SAQs before invitation and at screening six weeks later. RR = 89.5% (baseline questionnaires completed)	Prior to invitation and at screening	Anxiety and depression scales were significantly lower at screening than at baseline (p<.002, p<.0001). At screening, 19.9% obtained a clinically significant anxiety score while 5.7% obtained a clinically significant depression score. Some women reported stress-related behavior changes in the week before screening. Sleep, worry, and the ability to concentrate, relax, and feel happy were adversely affected and subjects reported more irritability. However, the proportion of women who reported these changes was modest (from 7% for ability to feel happy to 17.8% for sleeping). Adverse changes were correlated with anxiety and depression.

RR = response rate; * = response rate calculated by reviewers; SAQ = self-administered questionnaire.

Conclusions

The agenda for the study of abnormal mammograms should include the following areas.

1. Research is needed to characterize the impact of abnormal mammograms. Answers to the following questions are needed.
    - What are the psychosocial consequences of abnormal mammography and how long do they last?
    - How does abnormal mammography affect adherence to subsequent mammograms?
    - Are the effects related to the index of suspicion?
    - What is the cumulative impact of more than one abnormal mammogram?
    - Who are the women at most risk for extreme distress?
    - Do women with a family history and/or an identified genetic mutation predisposing them to breast cancer need special attention?
    - What are the direct and indirect costs of abnormal mammograms?
2. Research is needed to improve communication throughout the mammography experience and especially for women with abnormal results.
3. Intervention research also is needed to develop and test cost-effective interventions to aid women in coping with abnormal mammograms.
4. Special interventions may be needed for women who experience extreme distress about the abnormal result.
5. The research should be methodologically rigorous, with adequate power and standardized measures and measurement points.

At present, there are more questions than answers. One of the more intriguing questions is why there has been so little inquiry in an area that is of such vital concern.

References

(1) Dawson DA, Thompson GB. Breast cancer risk factors and screening: United States, 1987. Vital Health Stat [10], 1990:172, 1–60.

(2) Breen N, Kessler L. Trends in cancer screening—United States, 1987 and 1992. MMWR Morb Mortal Wkly Rep 1995;45:57–61.

(3) Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR Am J Roengenol 1995;165:1373–7.

(4) Ellman R, Angeli N, Christians A, Moss S, Chamberlain J, Maguire P. Psychiatric morbidity associated with screening for breast cancer. Br J Cancer 1989;60:781–4.

(5) Lerman C, Rimer BK, Engstrom PF. Cancer risk notification: psychosocial and ethical implications. J Clin Oncol 1991;9:1275–82.

(6) Paskett E, Rimer BK. Psychosocial effects of abnormal Pap tests and mammograms: a review. J Womens Health 1995;4:73–82.

(7) Bartolucci G, Savron G, Fava GA, Grandi S, Trombini G, Orlandi C. Psychological reactions to thermography and mammography. Stress Med 1989;5:195–9.

(8) Bull AR, Campbell MJ. Assessment of the psychological impact of a breast screening programme. Br J Radiol 1991;64:510–5.

(9) Gram IT, Lund E, Slenker SE. Quality of life following a false positive mammogram. Br J Cancer 1990;62;1018–22.

(10) Gram IT, Slenker SE. Cancer anxiety and attitudes toward mammography among screening attenders, nonattenders, and women never invited. Am J Public Health 1992;82:249–51.

(11) Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med 1991;114:657–61.

(12) Lidbrink E, Levi L, Pettersson I, Rosendahl I, Rutqvist LE, de la Torre B, et al. Single-view screening mammography: psychological, endocrine and immunological effects of recalling for a complete three-view examination. Eur J Cancer 1995;31A:932–3.

(13) Sutton S, Saidi G, Bickler G, Hunter J. Does routine screening for breast cancer raise anxiety? Results from a three wave prospective study in England. J Epidemiol Community Health 1995;49:413–8.

(14) Swanson V, McIntosh IB, Power KG, Dobson H. The psychological effects of breast screening in terms of patients' perceived health anxieties. Br J Clin Pract 1996;50:129–35.

(15) Epstein SA, Lin TH, Audrain J, Stefanek M, Rimer B, Lerman C. Excessive breast self-examination among first-degree relatives of newly diagnosed breast cancer patients. High Risk Breast Cancer Consortium. Psychosomatics 1997;38:253–61.

(16) Lerman C, Trock B, Rimer BK, Jepson C, Brody D, Boyce A. Psychological side effects of breast cancer screening. Health Psychol 1991;10: 259–67.

(17) Baines CJ, To T, Wall C. Women's attitudes to screening after participation in the National Breast Screening Study. A questionnaire survey. Cancer 1990;65:1663–9.

(18) Fine MK, Rimer BK, Watts P. Women's responses to the mammography experience. J Am Board Fam Pract 1993;6:546–55.

(19) Valdimarsdottir HB, Bovbjerg DH, Kash KM, Holland JC, Osborne MP, Miller DG. Psychological distress in women with a familial risk of breast cancer. Psychooncology 1995;4:133–41.

(20) Walker LG, Cordiner CM, Gilbert FJ, Needham G, Deans HE, Affleck IR, et al. How distressing is attendance for routine breast screening? Psychooncology 1994;3:299–304.

(21) Lerman C, Ross E, Boyce A, Gorchov P, McLaughlin R, Rimer BK, et al. The impact of mailed psychoeducational materials to women with abnormal mammograms. Am J Public Health 1992;82:729–30.

(22) Scudds RJ, Pederson LL. Phone, paper and pencil, in person: Methods of data collection for the '90s. Am J Health Behav 1996;20:443–7.

(23) Lerman C, Daly M, Sands C, Balshem A, Lustbader E, Heggan T, et al. Mammography adherence and psychological distress among women at risk for breast cancer. J Natl Cancer Inst 1993;85:1074–80.

(24) Lerman C, Lustbader E, Rimer B, Daly M, Miller S, Sands C, et al. Effects of individualized breast cancer risk counseling: a randomised trial. J Natl Cancer Inst 1995;87:286–92.

(25) McCarthy BD, Yood MU, Boohaker EA, Ward RE, Rebner M, Johnson CC. Inadequate follow-up of abnormal mammograms. Am J Prev Med 1996;12:282–8.

(26) Cardenosa G, Eklund GW. Rate of compliance with recommendations for additional mammographic views and biopsies. Radiology 1991;181: 359–61.

(27) King E, Rimer BK, Seay J, Balshem A, Engstrom PF. Promoting mammography use through progressive interventions: is it effective? Am J Public Health 1994;84:104–6.

Notes

Support for preparation of this manuscript was provided by NCI funded grants 1R01CA63782 and 1PO1CA72099.

Thanks are due to Steven Lewis for assistance with table preparation and Shirley Howard for word processing.

Variation of Benefits and Harms of Breast Cancer Screening With Age

*Russell Harris**

The critical issue in deciding whether to recommend breast cancer screening for women in their forties is to determine whether potential benefits are substantially greater than potential harms. Recent evidence from randomized clinical trials makes it likely that, after 10–12 years of follow-up, there is a real benefit from screening women ages 40–49, on the order of a 15–20% reduction in the relative risk of breast cancer death. This relative risk reduction translates into an absolute risk reduction of 1–2 women whose lives are extended from screening 1,000 women in their forties annually for 10 years (i.e., about one life extended per 5,000 mammograms). The absolute benefit of screening increases with age. Evidence about potential harms is less well established, but it is compelling that there are 15–40 times as many false positive as true positive mammograms (depending on the patient's age), and that at least some of the women with false positive mammograms have ongoing psychological distress as a result. Some 30% of all women who are screened annually during their forties will have at least one false positive mammogram and this probability likely decreases with advancing age. If the balance between benefits and harms is judged to be a "close call" for women in their forties, a blanket recommendation for all is inappropriate. Instead, each woman in her forties should be helped to understand the pros and cons of screening, to clarify her own values, and to consider with her primary care physician what decision would be best for her. [Monogr Natl Cancer Inst 1997;22: 139–143]

Getting the Question Right

The question I wish to address is what level of recommendation to make to women of different ages about breast cancer screening. I want to emphasize the phrase "what level of recommendation." Some may think the answer is a simple "yes" or "no"—either we recommend or we don't. The strength of the recommendation, however, should depend on the strength of the evidence about two issues: the benefits of screening *and* the harms of screening. The real question, then, is not whether there is some small benefit demonstrated for screening women in their forties. The issue is larger than a "*P*-value." What we need to know is where the balance lies between the magnitude of benefits and harms for different age (or other risk) groups.

But what do we do in cases where the balance between benefits and harms is not clear, as I believe is the case with breast cancer screening for women in their forties? In these cases, there is a third option beyond recommending or not recommending. Physicians may also raise the issue of breast cancer screening with their patients, help them understand the benefits and harms, and encourage them to participate in making an individualized decision. My aim, then, is to provide an overview of the benefits and harms of screening for women in their forties, so that these women, with the help of their physicians, can make the most appropriate decision for themselves.

Mortality Benefits of Screening

Screening seeks to decrease the risk of dying of breast cancer, not the risk of getting it. The specific risk a woman is trying to reduce by being screened for the next 10 years is the risk of eventually dying of cancer diagnosed in those next 10 years. These risks for women of different ages, calculated from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) data before widespread screening, are given in the second column of Table 1 (*1*). Not surprisingly, the risk increases with age.

To date, eight randomized trials of mammography screening among women aged 40 and older have been conducted in Sweden, the United Kingdom, Canada, and the United States. Mortality reduction in these trials is measured in terms of a "relative risk" reduction—that is, the reduction in risk of dying of breast cancer in a screened group relative to the baseline risk in an unscreened group. When the relative risk reduction from these randomized trials (Table 1, column 3) is factored in, we can calculate the absolute risk reduction (Table 1, column 4)—the number of women per 1,000 whose lives would ultimately be extended by screening over 10 years. The new evidence from the Swedish randomized trials makes it likely that there is a real benefit, and that it is on the order of a 15–20% relative risk reduction. If the relative risk reduction for women in their forties is 10–15%, then one woman would have her life extended for screening 1,000 women for 10 years. If the relative reduction for this age group is about 20%, then the lives of about two women would be extended for screening 1,000 women for 10 years (about one life extended per 5,000 mammograms).

Table 1 illustrates that the benefit of screening—the number of women per 1,000 whose lives are extended—increases with age. Some have made the claim that the benefit of screening is the same for women in their forties as for those in their fifties or sixties. These claims have used the relative risk reduction as a measure of benefit. It is clear from Table 1, however, that even

**Affiliation of author:* Division of General Medicine and Clinical Epidemiology, Department of Medicine, UNC Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC.

Correspondence to: Russell Harris, MD, MPH, CB# 7508, Building 52, UNC School of Medicine, Chapel Hill, NC 27599–7508.

Table 1. Benefits of screening

Age	Risk per 1,000 women*	Relative risk reduction (%)	Absolute risk reduction†
40	7.8	16**	1.2
		23††	1.8
50	12.9	15	1.9
		30	3.9
60	19.5	30	5.9
≥70	25.3	30‡‡	7.6

*Rate of dying in next 15–20 years of breast cancer diagnosed in next 10 years, from SEER data, 1973–1980 and 1989–1991.
†Number of lives ultimately extended per 1,000 by screening over the next 10 years.
**From Swedish meta-analysis.
††From Edinburgh trial, beginning with age 45 years.
‡‡Extrapolated from 60–69 age group.

if the relative risk reduction is equivalent in different age groups (which is not at all certain from the trials), the absolute benefit in number of lives extended per 1,000 women screened increases with age.

Effects of Screening on Nonmortality Outcomes

Screening is a ''double-edged sword'' that can result in either benefits or harms. There is a need for more research on both nonmortality benefits and harms of screening. The potential magnitude of the effects, however, is apparent from examining the screening ''cascade''—that is, the expected sequence of events following screening. This cascade is shown in Figs. 1 and 2 for a single screening of a hypothetical population of 10,000 women who are being screened regularly (i.e., ''incidence'' screens as opposed to ''prevalence'' screens). The figures assume a conservative mammogram ''positivity'' rate of 5%—that is, 5% of all cases will require further evaluation. This rate

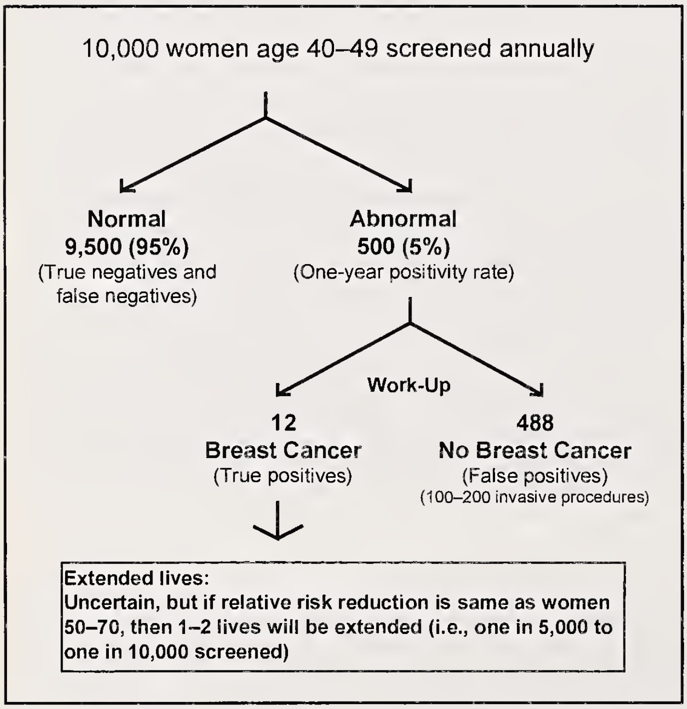

Fig. 1. Extended lives: 2–6 each year (one in 1,700 to one in 5,000 screened). Reprinted with permission from (*1*).

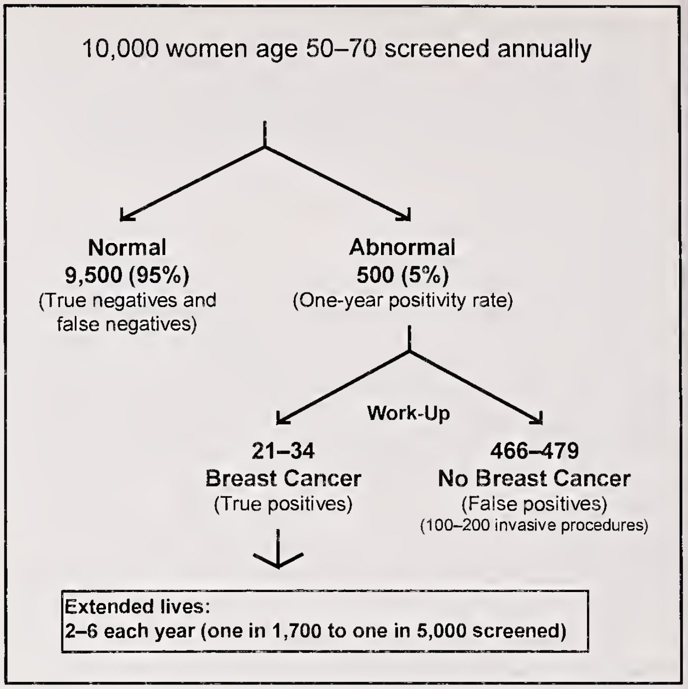

Fig. 2. Extended lives: uncertain, but if relative risk reduction is same as women 50–70, then 1–2 lives will be extended (i.e., one in 5,000 to one in 10,000 screened). Reprinted with permission from (*1*).

is indicative of many excellent mammography practices (*2*) and is much less than the 11% found in a recent national survey (*3*). Using the higher rate would double the number of false positives. Sensitivity of mammography is taken from average sensitivity in the trials (*4*), and incidence of cancer is taken from SEER data (*5*).

Screen-Negative Women

As seen in both figures, most women screened are negative. The great majority of these women are truly negative—they do not have breast cancer. A few, however, truly have cancer but are screen negative—that is, they are falsely negative. A research priority is to find out whether some of these women have been injured by false reassurance. It seems possible that some may ignore early symptoms of breast cancer because they have been reassured by the negative mammogram. We don't know how often this really happens.

The true negatives would seem to be in a position to benefit; they could receive ''peace of mind''—reassurance that they do not have cancer. But if you look carefully at the probability of having cancer before screening versus after negative screening—which for women in their forties is about 1.6 per 1,000 before screening and about 0.4 per 1,000 after a negative screen, a reduction in risk of about 1 per 1,000)—the difference doesn't seem large enough to make a truly objective woman change from worrying to relaxing. The woman was at low risk before screening and is still at low (but not zero) risk after being screen negative. Nevertheless, many women report peace of mind after a negative mammogram. This may reflect overestimation of initial risk and overinterpretation of a negative mammogram, and it suggests that we should develop ways other than mammography of reassuring women.

Screen-Positive Women

In many excellent mammography practices, about 5% of women are screen positive, and the great majority of these are falsely positive (i.e., they do not have cancer, despite the positive test). As shown in Figs. 1 and 2, there are 15–40 times as many false positives as true positives. And Figs. 1 and 2 are only for a single screen. The cumulative probability of having at least one false positive over 10 years of screening is unknown and should be a research priority. This probability could easily be as high as 30% (*6*) (or more) of all women.

All screen-positive women subsequently undergo a "work-up," which may be fine needle aspirate, ultrasound, or magnification views. Some will come to biopsy. We are only now beginning to appreciate the experience of women who face the burden of a false positive mammogram. Although more research is needed, it is clear now that many of these women will have marked anxiety in the days (sometimes weeks) between learning of their abnormal mammogram and being told that they do not have cancer. Some of these women will have continued anxiety months after being told that they do not have breast cancer (*7*). The experience of this large group of women should be a prime consideration in deciding whether to recommend screening.

A related research priority should be to find ways to minimize the psychological trauma for false positive women. It is incorrect, however, to assume that this trauma can be erased completely by various interventions. It is entirely possible that at least some of this anxiety is inherent in the screening situation and in our current societal views of breast cancer.

Women who are "true positive"—those who screen positive and are found to actually have breast cancer—are the women we usually think have been helped most by screening. Unfortunately, not all women whose cancer is detected by screening benefit from that detection. Breast cancer is a heterogeneous disease with a spectrum of natural histories (*8*). For our purposes here, we can simplify this spectrum into three distinct types. About 50% of true-positive women will not die from breast cancer, even if they are never screened and wait until later in life for their cancers to be detected. These cancers are slow growing and relatively treatable. Screening will not alter their natural history because their natural history is excellent. Still, the perception of many of these women, quite understandably, is that their lives have been "saved" by screening.

Another type of breast cancer has an aggressive natural history and is difficult to treat. Women with this type of tumor, unfortunately, will die of breast cancer regardless of when it is found. These cancers metastasize at an early, undetectable stage. Again, the natural history of the disease is not altered by screening, and hence, there is no benefit to screening. The woman who has had an aggressive cancer found by screening has simply been made to live longer with knowledge of the diagnosis. One could argue that these women have been harmed, not helped, by screening.

Finally, some cancers are more treatable when found earlier, and thus screening favorably changes their natural history. The screening trials help us estimate the number of women with this type of cancer. As shown in Table 1, the randomized screening trials indicate that somewhere between 10% and 25% of women who would have died of breast cancer have this type of cancer, i.e., that is more treatable if found early. The most recent Swedish data narrow this estimate to 15%–20%. This, then, translates into 1–2 lives extended per 10,000 women screened once (or, as noted above, 1,000 women screened annually for 10 years), or about one life extended per 5,000 mammograms.

Ductal Carcinoma *In Situ*

In addition to women who are true positive for invasive breast cancer, some will be found to have ductal carcinoma *in situ* (DCIS). The natural history of DCIS is unknown. Some, but likely not all, of these lesions will progress to invasive carcinoma. And when progression occurs, it may take many years (thus allowing opportunities for detection at a later age) (*9*). Understanding the natural history of DCIS and determining the characteristics of those lesions that will become clinically important as opposed to those that are actually "pseudodisease" (a pathologic finding that never produces clinical disease) should be a research priority. If, as we suspect, 50% or more of these lesions are clinically unimportant, then the potential for harming women by unnecessary treatment could be an important factor for women to consider in deciding about screening.

Less Intensive Treatment

One potential benefit of screening is the possibility that women whose cancers are found at an earlier stage will require less intensive therapy. Unfortunately, there are insufficient data to determine whether this theoretical benefit is real. Certainly many women with palpable tumors (not found by screening) are still eligible for lumpectomy rather than mastectomy. And many small, node-negative tumors (as well as DCIS) are being treated with surgery and either radiation or adjuvant chemotherapy. It is not clear whether increased screening has led to more or less intensive therapy for the population as a whole.

Variation of Harms by Age

As shown in Table 2, some of the potential harms of breast cancer screening vary with age. Because the sensitivity of mammography is lower for younger than older women, yet there are more total cancers among older women, the number of false negatives is similar in the different age groups. True positives are more frequent in older women, although for all women the number of true positives is small relative to false positives. The incidence of DCIS increases gradually with age, and thus we can expect that there will be slightly more women with this lesion in their fifties and sixties as compared with women in their forties.

Table 2. Harms of screening

Type of finding	Harm	Relationship with age
False-negative	False reassurance	40 ≅ 50/60*
False-positive	Psychological trauma	40 > 50/60
True-positive No change in natural history	Living longer with knowledge of disease	50/60 > 40
Pseudodisease Ductal carcinoma	Labelling-psychological effects Unnecessary treatment	50/60 > 40

*Rate for women in their forties compared to rate for women in their fifties or sixties. (40 = women in their forties; 50/60 = women in their fifties and sixties).

By far the largest group of women who may be harmed by screening is the false positive group. As noted earlier, this group may include as many as 30% of women in their forties screened annually for 10 years. A critical question, then, is whether the probability of a woman becoming a false positive varies by age. An important determinant of the probability of having a false positive is the initial "positivity rate" of screening—that is, the percentage of women screened who required some further work-up. There is conflicting evidence about whether this percentage varies with age. In some studies, especially those of academic practices (*10*), the positivity rate appears fairly constant with age. In studies of community practices (unpublished data, New Hanover Breast Cancer Screening Study, 1990; personal communication, Nancy Lee, M.D., from National Breast and Cervical Cancer Early Detection Program; personal communication, Bruce McCarthy, M.D.), younger women have higher positivity rates (and thus more false positives) than older women (*2*). The issue is important and should be a research priority. Even with the same positivity rate, however, the fact that the incidence of breast cancer is higher in older women means that more of the positives in younger women will be falsely positive.

But there is another factor that makes it very likely that women in their forties have a larger—even a much larger—probability of a false positive than older women. This other factor is the frequency of screening. From the trials of women over 50, it appears that a large percentage of the benefit of annual screening can be obtained by screening biennially. For women in their forties, however, it is clear that if screening works at all, it must be done annually. Although there is need for research in this area, it seems likely that screening twice as frequently would produce a higher cumulative rate of false positive findings than screening biennially.

The bottom line is that breast cancer screening is not the final answer to the problem of breast cancer in any age group. It certainly has benefits, however, among women ages 50 to 70 years, and, as shown by the Swedish studies, probably benefits as well for women in their forties. The benefit for women in their forties is delayed and small in terms of absolute number of lives extended per 1,000 women screened. Benefits gradually increase with age, and harms, flowing largely from the number of false positives, gradually decrease with age.

Restating the Problem

The problem, then, can be immediately appreciated. As they grow older, even well-informed women will naturally differ in their perceptions of the age at which the increasing probability of benefit outweighs the decreasing probability of harm. And policy makers will naturally differ in their evaluation of the age, on the population level, at which the increasing benefits of screening begin to outweigh the decreasing harms. Perhaps the disagreement should tell us something. We differ not because we disagree about what the evidence is, but rather because our values differ. There is no consensus about screening for women in their forties, nor should there be. This is a "close-call." In such situations, women should be helped to participate in their own decisions.

Some may question whether it is feasible for medical practices to help women understand the potential benefits and harms of screening, and to facilitate informed, shared decision making. Finding time for such discussions may be difficult in busy medical practices. A research priority should be to develop and evaluate "discussion aids," such as videotapes, decision boards, and tailored brochures, as well as training of nonphysician staff, to help medical practices accomplish this task more efficiently and effectively.

Some may also question whether many women will want to participate in such a decision, when their physicians do not make a strong recommendation. But the reason for not making a recommendation is not lack of information; it is rather the understanding of the issue as a "close call." In such situations, women's values and perceptions will carry as much weight as the facts about the pros and cons of screening. Ideally, the patient herself should supply such information to the decision-making process. We need to better understand how women will react when encouraged to participate with their physicians (and other members of the medical staff) in a process of informed, shared decision making.

Population Level

This analysis has focused on the individual level, and the need to help individual women to participate in a process of individualized, informed decision making. One could also take a population perspective. Because of the relatively low risk of a woman dying of breast cancer diagnosed in her forties, a risk reduction of 15%–20% turns out to be a small absolute risk reduction for an individual. However, this number becomes larger if the 15%–20% is multiplied times the total number of women dying each year of breast cancer diagnosed during their forties. If, for example, about 5,000 women each year die of breast cancer diagnosed during their forties, then screening could conceivably extend the lives of 750–1,000 women each year (assuming 100% compliance with screening). Unfortunately, the harms (not to speak of the financial costs) are also multiplied. Many more than 1,000 women would face the trauma of a false positive mammogram; many would have a biopsy; many would be diagnosed with DCIS. Again, this appears to be a close call, even on the population level.

Whether the decision is considered on the individual or population level, we should all be concerned by the lack of understanding of breast cancer risk and breast cancer screening by many American women. Several years ago, some colleagues and I surveyed women living in two eastern North Carolina counties, and found that worry about breast cancer was higher among women in their forties than women in their fifties and sixties. Less than 25% of women of any age understood that breast cancer risk increases with age (*11*). More recent surveys of nearly 4,000 women visiting primary care physicians found that over half of women in their forties overestimated their risk of breast cancer by a factor of three or more, and nearly half overestimated the benefit of screening mammography by a factor of at least 10 (unpublished data, North Carolina Prescribe for Health study, 1994). Others have found similar results (*12*).

A blanket recommendation that all women in their forties be screened would not serve the cause of public or individual education about this issue. Furthermore, such a recommendation

would be discordant with the weakness of the evidence that benefits outweigh harms. A more measured approach is needed. The recommendation for women in their forties should be that they be informed that there are pros and cons to being screened, and that reasonable women will disagree about whether to be screened. They should be encouraged to clarify their own values and then to discuss screening with their physicians, to participate in making an individualized decision. Then we should get to work on the real issue: how to efficiently and effectively reach all women with this discussion.

References

(*1*) Harris R, Leininger L. Clinical strategies for breast cancer screening: weighing and using the evidence. Ann Intern Med 1995;122:539–47.

(*2*) CDC. Results from the National Breast and Cervical Cancer Early Detection Program, October 31, 1991-September 30, 1993. MMWR Morb Mortal Wkly Rep 1994;43:530–4.

(*3*) Brown ML, Houn F, Sickles EA, Kessler LG. Screening mammography in community practice: positive predictive value of abnormal findings and yield of follow-up diagnostic procedures. AJR Am J Roentgenol 1995;165: 1373–7.

(*4*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.

(*5*) Ries LA, Miller BA, Hankey BF, Kosary CL, Harras A, Edwards BK, editors. SEER Cancer Statistics Review, 1973–1991: Tables and Graphs, National Cancer Institute. NIH Pub. No. 94-2789. Bethesda (MD), 1994.

(*6*) Elmore JG, Barton MB, Moceri VM, Fletcher SW. Cumulative risk of a false-positive mammogram over a 10-year period [abstract]. J Gen Intern Med 1997;12 Suppl:107.

(*7*) Lerman C, Trock B, Rimer BK, Boyce A, Jepson C, Engstrom PF. Psychological and behavioral implications of abnormal mammograms. Ann Intern Med 1991;114:657–61.

(*8*) Harris JR, Hellman S. Natural History of Breast Cancer. In: Diseases of the Breast. Harris JR, Lippman ME, Morrow M, Hellman S, editors. Philadelphia: Lippincott-Raven, 1996.

(*9*) Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996; 275:913–8.

(*10*) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.

(*11*) Harris RP, Fletcher SW, Gonzalez JJ, Lannin DR, Degnan D, Earp JA, et al. Mammography and age: are we targeting the wrong women? A community survey of women and physicians. Cancer 1991;67:2010–4.

(*12*) Black WC, Nease RF Jr, Tosteson AN. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. J Natl Cancer Inst 1995;87:720–31.

Nonpalpable Breast Cancer in Women Aged 40–49 Years: A Surgeon's View of Benefits From Screening Mammography

*Helena R. Chang, Bernard Cole, Kirby I. Bland**

While mammography screening among women aged 50 years or older has proven to reduce breast cancer mortality, screening in younger women has been repeatedly scrutinized. To test the effect of screening among younger women, we examined 84 consecutive patients aged 40–49 at the time of breast cancer diagnosis: 27 (32.1%) were diagnosed solely by mammography, and 57 (67.9%) had a palpable mass. The mean tumor sizes were 1.3 cm and 3.6 cm for the two groups respectively. While 68.8% nonpalpable invasive tumors were classified as Stage I cancer, only 34% of patients with palpable breast cancer had Stage I disease. None of the patients with nonpalpable breast cancer had disease beyond Stage II. In contrast, 28.3% of the patients with palpable invasive breast cancer presented with advanced disease. In addition, 6.3% versus 41.5% of patients with nonpalpable and palpable breast cancer respectively had nodal metastases. The five-year survival rates for the two groups were 100% and 73% respectively, favoring breast cancer detected mammographically. Screening of women aged 40–49 also resulted in more breast-conserving surgery and less chemotherapy. We conclude that screening in this age group should be continued, although individual assessment is needed. [Monogr Natl Cancer Inst 1997:22:145–149]

The beneficial effect of screening mammography among women aged 50 and older has been consistently demonstrated worldwide (*1–6*). Indeed, screening has been recognized as the most effective tool against breast cancer in this age group, and it has been firmly recommended for all women aged 50 and older. Meanwhile, the debate regarding its usefulness for women aged 40–49 remains unsettled (*7–13*). Recently, however, a meta-analysis of seven randomized trials studying women aged 40–49 years has demonstrated a statistically significant 24% reduction in breast cancer mortality due to screening intervention (*14*). It has been suggested that this outcome may be further improved by annual two-view screening with high-resolution mammography (*15,16*).

These results, together with the significant breast cancer incidence in young women and the subsequent loss of life, make any negative recommendation for screening an extremely serious public health concern. It is estimated that a 40-year-old woman has a 1 in 63 chance of developing breast cancer before age 50 (*17*). Approximately 18% of all breast cancers (*18*), 20% of all breast cancer deaths, and one-third of all years of life expectancy lost due to breast cancer are in women of this age group (*5*). Any guidelines recommended by health professionals should therefore target improvements in the diagnosis, survival, and quality of life after diagnosis for this age group.

While all the randomized trials to date have focused on the reduction of cancer death by screening programs, each has overlooked important quality-of-life issues, such as whether a woman must undergo adjuvant therapy and whether breast-conserving surgery is a treatment option. The purpose of this paper, therefore, is to evaluate not only mortality benefits due to screening, but also subsequent improvements in quality of life. Specifically, we compare tumor size, cancer staging, surgical treatment, adjuvant therapy, and disease control between women aged 40–49 years with palpable tumors and women of this age with nonpalpable breast cancer detected by mammography.

Patients and Materials

Eighty-seven breast cancer patients aged 40–49 were identified in a single institution between 1983 and 1995. Patients with mammographically detected nonpalpable breast cancer were identified by reviewing the operative notes, specimen mammography, and pathology reports. Specimen mammography was performed after tissue was removed by the hook-wire method to ensure inclusion of the concerned area. When patients were either biopsied or surgically treated elsewhere, the pathologic confirmation of breast cancer diagnosis was achieved by institutional review. In these cases, the palpability of the original tumor was determined from the treating physician's notes and categorized as either nonpalpable (removed by needle localization) or palpable. The palpability of one tumor was uncertain, and it was excluded from the analysis. Since this study was aimed at examining the value of screening mammography, two cases with nonpalpable but nonmammographically detected breast cancer were also excluded: one patient had Paget's disease of the nipple, and the other patient had an incidental finding of ductal carcinoma *in situ* (DCIS) and lobular carcinoma *in situ* (LCIS) from breast-reduction specimen.

Mean size of invasive primary cancer, cancer staging, nodal

**Affiliations of authors:* H. R. Chang, Department of Surgery, Roger Williams Medical Center, Brown University, Providence, Rhode Island; B. Cole, Center for Statistical Sciences, Brown University, Providence, Rhode Island; K. I. Bland, Rhode Island Hospital and Department of Surgery, Roger Williams Medical Center, Brown University, Providence, Rhode Island.

Correspondence to: Helena R. Chang, M.D., Ph.D., Department of Surgery, Roger Williams Medical Center, Brown University, 825 Chalkstone Avenue, Providence, RI 02908.

See "Note" following "References."

status, types of surgical treatment, need for adjuvant treatment, overall survival, and disease-free survival were compared in the two groups. Tumor size was defined as the maximal diameter of the gross lesion or the microscopic measurement of the nonapparent lesion. The choice of surgical treatment was jointly decided by the treating surgeon, the radiation oncologist, and the patient. When needed, adjuvant therapy was recommended by a multidisciplinary team at the institution after team members reviewed the complete pathologic findings of the primary breast cancer and axillary lymph nodes.

The statistical significance of differences of all parameters between the two groups was analyzed by Fisher's exact test or by chi-square analysis. Survival time was measured from the date of diagnosis. The survival curves were generated using the Kaplan-Meier method, and the survival curves were compared by the log-rank test.

Results

Eighty-four women aged 40–49 years with breast cancer were identified between 1983 and 1995. Eighty-two percent were found to have invasive cancer, and the remaining had *in situ* disease. Approximately one-third of young patients (n = 27) had mammographically detected breast cancers, which were surgically removed by needle-guided breast biopsy. Of these 27 patients, 40.7% were found to have *in situ* breast cancer. In contrast, only 5.2% of patients with palpable breast cancer had the same premalignant condition. A palpable mass was strongly associated with invasive breast cancer (Table 1).

The size of the primary invasive cancer was compared between mammographically detected cancers and those diagnosed palpably. The mean size of the tumors in the mammographically diagnosed group was 1.3 cm, with a median tumor diameter of 0.8 cm. This was much smaller than the mean tumor diameter of 3.6 cm ($P = 0.059$) and the median diameter of 2.5 cm ($P = 0.003$) in patients with palpable cancers (Table 2). Forty-four percent of the nonpalpable breast cancers were 1 cm or less, compared to 16% of palpable breast cancers. The difference in distribution of tumor size in the two groups was significant ($P = 0.049$), with the group having nonpalpable tumors dominated by small cancers (Table 3). Five of the palpable cancers had no definitive size due to diffuse involvement of the breast or metastatic disease. Tumor diameter was not available in three patients with nonpalpable breast cancer, all of whom had either malignant microcalcifications or fragmented specimens, and hence the exact sizes could not be correctly calculated.

In addition to being smaller tumors, mammographically detected cancers also tended to be early-stage cancers. More than two-thirds of the mammographically detected cases had Stage I disease, and none had disease beyond Stage II. In contrast, only one-third of the patients with palpable tumors had Stage I breast cancer, and approximately one-third had Stage III and Stage IV breast cancer (Fig. 1). The difference in stage distribution between the two groups of patients was statistically significant ($P<0.001$). The incidence of nodal metastasis in the two groups also differed significantly, with 6.3% in the former group and 41.5% in the latter group ($P = 0.046$) (Table 4). The mean numbers of metastatic nodes were 3.34 for patients with palpable breast cancer and 0.05 for patients with nonpalpable breast cancer ($P = 0.0496$).

Breast conservation was the most common form of surgical treatment for mammographieally discovered breast cancer (Table 5). The mean tumor size was 1.4 cm for those who received lumpectomy and 1.3 cm for patients who received mastectomy. It is possible that all these patients were candidates for breast-conserving surgery. In comparison, the majority of patients with palpable breast cancer were treated with mastectomy. The mean tumor size of those who received a mastectomy for a palpable cancer was 4 cm, which appeared to justify the choice of mastectomy.

Postoperative chemotherapy was less frequently employed in treating patients with mammographically detected breast cancer. While 67% of the women with palpable invasive breast cancer had chemotherapy, only 31% of the group with nonpalpable breast cancer received multidrug chemotherapy ($P = 0.01$). More conservative surgery and less chemotherapy did not pose any adverse effect on the excellent outcome of patients with

Table 1. Comparison of characteristics of women aged 40–49 years with palpable versus mammographically detected (MD) nonpalpable breast cancer in 84 patients (1983–1995)

Characteristics	Breast cancer: Palpable	Breast cancer: MD (nonpalpable)†
Mean age	45 years	46 years
Race		
White	55	22
Nonwhite	0	1
Unknown	3	4
Invasive cancer	53	16
Ductal	42	10
Lobular	1	2
Ductal and lobular	0	1
Other	10	3
DCIS	4	11
Total cases	57	27

†Nonpalpable breast cancer diagnosed by needle-localization–guided breast biopsy.

Table 2. Mean size of palpable vs. mammographically detected (MD) nonpalpable invasive breast cancer in women aged 40–49 years

Tumor size	Invasive breast cancer: Palpable	Invasive breast cancer: MD	Statistical difference
Mean	3.6 cm	1.3 cm	$P = 0.059$
Median	2.5 cm	0.8 cm	$P = 0.003$

Table 3. Distribution of tumor size in women aged 40–49 with palpable vs. mammographically detected invasive breast cancer

Tumor size	Breast cancer: Palpable	Breast cancer: MD
≤1 cm (T1a,b)	8 (15.0%)	7 (43.8%)
1.1–2.0 cm (T1c)	15 (28.3%)	3 (25.1%)
2.1–5.0 cm (T2)	19 (35.8%)	3 (12.5%)
>5.0 cm (T3)	6 (11.3%)	0
Unknown	5* (9.4%)	3 (18.8%)

$P = 0.049$.

*Five of the palpable cancers had no definitive size due to diffuse involvement of the breast or metastatic.

Table 4. Axillary metastases in women aged 40–49 with palpable vs. mammographically detected invasive cancer

Nodal status	Breast cancer Palpable	MD
+	22 (41.5%)	1 (6.3%)
–	28 (52.8%)	14 (87.5%)
Unknown	3 (5.7%)	1 (6.3%)

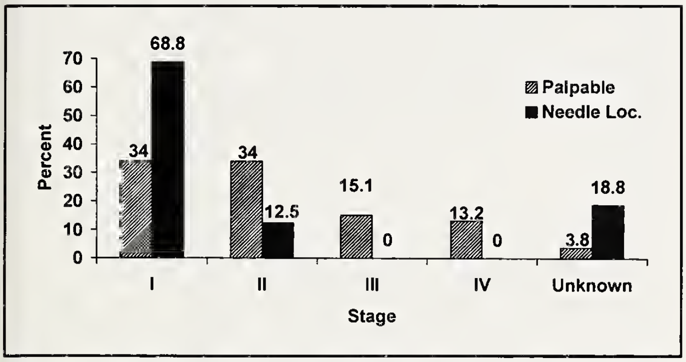

Fig. 1. Distribution of cancer stage in women aged 40–49 years with either palpable or mammographically detected (MD) nonpalpable invasive breast cancer (P<0.001).

Table 5. Type of surgery received by women aged 40–49 with either palpable or mammographically detected breast cancer

	Breast cancer Surgery palpable/mean tumor size	MD/mean tumor size
Lumpectomy	19 (33.4%)/2.6 cm	14 (51.8%)/1.4 cm
Mastectomy	33 (57.3%)/3.9 cm	12 (44.4%)/1.3 cm
Neither/unknown	5 (9.3%)	1 (3.8%)

mammographically discovered breast cancer. Their five-year overall survival and disease-free survival rates were both 100% (Fig. 2). In contrast, the five-year overall and disease-free survival rates were 70% and 62% respectively among patients with palpable breast cancer. However, the five-year survival rates associated with women with local disease, regional disease, and distant metastases in this latter group were 89.7%, 68.9%, and 17.9% respectively, suggesting that the survival rates were stage specific and were not adversely affected by tumor palpability alone.

A significant proportion of women aged 40–49 in the study had mammographically detected breast cancer. The breast cancer detected by this mode resulted in 40% of *in situ* disease. Among those patients aged 40–49 with invasive breast cancers, 94% were free of nodal metastasis. Mammographically detected cancers among these young women tend to be small in size and early in cancer stage. Young women with mammographically discovered breast cancer were more likely to receive breast conservation surgery and less likely to require chemotherapy. The excellent survival rate and disease control simply reflect that screening mammography detects breast cancer at a favorable stage.

Discussion

Breast cancer is the leading cause of death in women in their forties in the United States (*19*). Two screening methods—mammography and clinical breast examination—are thought to be life saving for women over 50 years of age, but the same techniques have been suggested by some to be ineffective for women aged 40–49 years. Therefore, women in the younger age group are not screened routinely and must often wait for breast cancer to appear clinically before being treated.

The opponents of universal screening for women in their forties have been supported by the report of the Canadian National Breast Screening Study (NBSS). The NBSS reported that more node-positive breast cancer cases and more patients with four or more positive lymph nodes were found in a mammographically

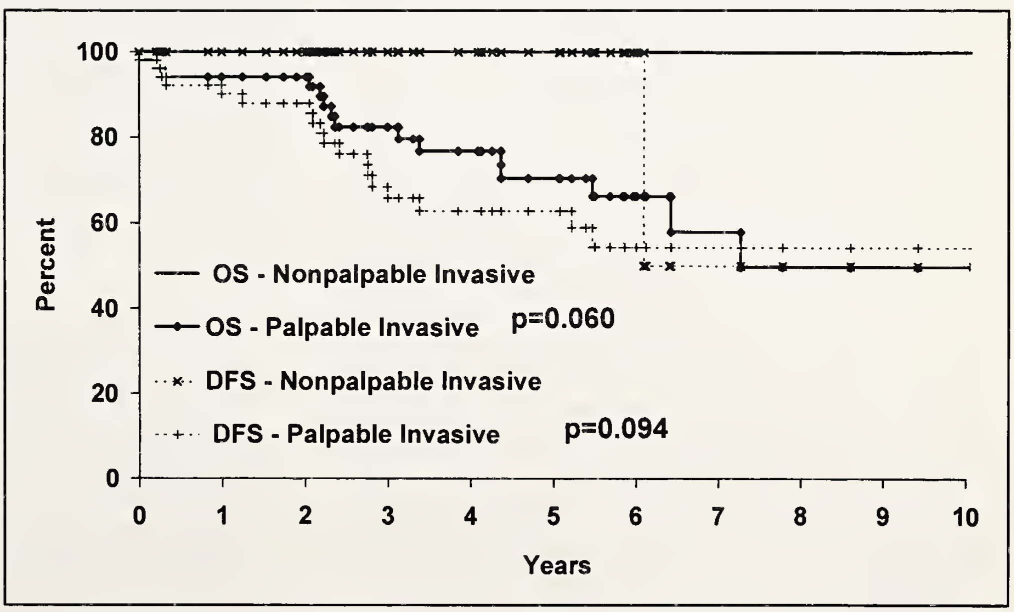

Fig. 2. Five- and ten-year overall survival rates in women aged 40–49 years with either palpable or mammographically detected (MD) nonpalpable invasive breast cancer. (P = 0.060).

screened group than in controls. This study implied that screening mammography caused more advanced breast cancer locally and regionally, hence a higher breast cancer mortality.

According to our findings, this implication is misleading and unfounded. Our study focused on the characterization of breast cancers that were detected by mammography in asymptomatic women aged 40–49 years. Approximately one-third of young women in our study had nonpalpable cancer. The majority of these nonpalpable breast cancers found by mammography were either *in situ* tumors or small invasive breast cancers. The mean size of invasive cancers detected as nonpalpable, mammographic abnormalities was 1.3 cm, and 94% of these patients had negative lymph nodes. The five-year survival rate was 100%, which is significantly better than 70% observed in patients with palpable breast cancer. Patients rarely had recurrent disease, which was reflected by an excellent disease-free survival rate at five years. Furthermore, only 31.3% received adjuvant chemotherapy. In contrast, 67% of patients with palpable breast cancer required ehemotherapy. None of the patients with nonpalpable breast cancer had either Stage III cancer or metastasis at the time of diagnosis. On the contrary, 14 of the 53 patients with palpable breast cancers were found to have advanced disease.

Our findings therefore support the cautious continuation of screening for women in their forties. This conclusion is supported by several previous studies, including of the Health Insurance Plan (HIP) trial, the first randomized controlled trial (RCT) of breast cancer screening. Although an initial, short follow-up of the HIP study reported no survival benefits to screening among women aged 40–49 years (*5,20*), an 18-year follow-up demonstrated a 24% reduction in the mortality of women who entered the study at ages 40–49 years (*21*). A second U.S. study, the Breast Cancer Detection and Demonstration Project (BCDDP), has been remarkable for demonstrating superior detection of breast cancer not only among postmenopausal women, but also women aged 40 to 49. In the BCDDP study, the breast cancers deteetion rate by screening mammography was 90% for young women and 92% for women aged 50–59 years. The improved mammographic capability resulted in detecting smaller breast cancer, and 80% of all breast cancers detected by screening mammography were free of nodal metastases (*22, 23*). The overall 14-year adjusted survival rate for women aged 40–49 years with invasive breast cancer was 81.8%.

In 1993, Fletcher reported that screening mammography was not beneficial for women aged 40–49 years based on a follow-up of five to nine years from previously conducted RCTs (*24*). More recently, however, an analysis of the five Swedish trials, which included 282,777 women who were followed for 5 to 13 years, showed a 13% reduction of breast cancer mortality among screened women aged 40–49 years (*25*), although the beneficial effect did not emerge until after eight years of follow-up. The Edinburgh trial also revealed no benefit for the first seven years of follow-up. However, the relative risk (RR)—the breast cancer mortality rate of screened women relative to nonscreened women—decreased significantly at the 10-year follow-up (*26*). Kerlikowske *et al.* notes that in those clinical trials in which women aged 40–49 years underwent two-view mammography and had 10 to 12 years of follow-up, the RR of screened to nonscreened women decreased significantly (RR = 0.73) (*27*). Another meta-analysis of the seven prospective randomized trials included women aged 40–49 years, reporting a 24% reduction of breast cancer mortality by breast cancer screening (*14*). Taken together, then, these studies suggest a beneficial effeet from screening, at least after about 8 to 10 years.

In addition to suggesting a benefit from screening mammography for young women, analysis of the National Cancer Institute's Surveillance, Epidemiology, and End Result Program data showed that the breast cancers detected solely by mammography were mainly DCIS and lesions less than 1.9 cm in diameter with no axillary nodal involvement (*30*). The BCDDP study further showed that the rates of breast cancer detected by mammography in women aged 40–49 and women aged 50–59 were similar. Kopans *et al.* found that the positive predictive value of breast biopsy performed as a result of mammography does not abruptly change at age 50 years (*29*). Thus, it is inappropriate to assume that screening mammography in women aged 40–49 years is ineffective.

Studies directed to examine the outcome of needle-localized breast biopsies in women aged 40–49 are few. Lein *et al.* (*30*) recently examined 207 patients in this age category who underwent needle-guided biopsies. Fifteen percent of these patients were found to have breast cancer. Although the mean tumor diameter of this particular age group was not specified, the mean tumor size of all age groups was 1.46 cm. Others found that a high percentage of young women with occult, grouped microcalcifications had early-stage or noninvasive ductal carcinoma detected by screening mammography (*31,32*). Wilhelm *et al.* demonstrated that patients with nonpalpable breast cancer detected by mammography tended to have small tumors, fewer nodal metastases, and better survival (*33–35*). Even the NBSS study pointed out that the best survival was found in young women whose breast cancers could only be detected by mammography.

Aside from favorable size, nodal status, and survival outcomes of patients with mammographically detected breast cancer, Haffty *et al.* has demonstrated excellent local control and overall survival in patients with nonpalpable breast cancer who are treated by breast-conserving surgery and radiation. These authors further point out that none of these patients received adjuvant chemotherapy (*36*). One would assume that the quality of the patient's life is likely to be improved when breast-conserving treatment is an option, and when it is possible to avoid chemotherapy among women with mammographically detected breast cancer.

Based on our study and the work of others, then, we believe that there is insufficient evidence to discontinue the practice of screening mammography in women aged 40–49 years.

References

(*1*) Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. Radiol Clin North Am 1992;30:187–210.

(*2*) Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ 1988;297:943–8.

(*3*) Roberts MM, Alexander FE, Anderson I, Chetly U, Donnan PT, Forrest P, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet 1990;335:241–6.

(*4*) Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist LE, Somell A. Randomized study of mammographic screening—preliminary report on mortality in the Stockholm trial. Breast Cancer Res Treat 1991;18:49–56.

(*5*) Shapiro S, Venet W, Strax P, Venet L. Periodic screening for breast cancer:

The Health Insurance Plan project and its sequelae; 1963–1986. Baltimore (MD): Johns Hopkins University Press, 1988.
(*6*) O'Maley M, Fletcher S, Morrison B. Does screening for breast cancer save lives? Effectiveness of treatment after breast cancer detection following screening by clinical breast examination. In: Mammography and Breast Self-Examination. New York: Springer Verlag, 1990.
(*7*) Miller AB. Is routine mammography screening appropriate for women 40–49 years of age? Am J Prev Med 1991;7:55–62.
(*8*) Rutqvist LE, Miller AB, Andersson I, et al. Reduced breast cancer mortality with mammography screening: an assessment of currently available data. Int J Cancer 1992;5 Suppl:76–84.
(*9*) Tabar L, Dean PB. Breast Cancer Screening [letter]. Med J Aust 1991; 154:853–4.
(*10*) Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Report on a workshop of the UICC project on evaluation of screening for cancer. Int J Cancer 1990;46:761–9.
(*11*) Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years [published erratum appears in Can Med Assoc J 1993;148:718]. Can Med Assoc J 1992;147:1459–76.
(*12*) Nystrom L, Rutqvist LE, Wall S, Tabar G. Breast cancer screening for breast cancer. J Natl Cancer Inst 1993;85:1644–56.
(*13*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.
(*14*) Smart CR, Hendrick RE, Rutledge JH III, Smith RA. Benefit of mammography screening in women ages 40 to 49 years. Current evidence from randomized controlled trials [published erratum appears in Cancer 1995; 75:2788]. Cancer 1995;75:1619–26.
(*15*) Feig SA. Estimation of currently attainable benefit from mammographic screening of women aged 40–49 years. Cancer 1995;75:2412–9.
(*16*) Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer 1995;75:2507–17.
(*17*) Feuer EJ, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing cancer. J Natl Cancer Inst 1993;85:892–7.
(*18*) Mettlin C. The relationship of breast cancer epidemiology to screening recommendations. Cancer 1994;74 Suppl:221–30.
(*19*) Department of Health and Human Services. Vital Statistics of the United States. Cancer Mortality 1988:988.
(*20*) Shapiro S, Strax P, Venet L. Evaluation of periodic breast cancer screening with mammography: Methodology and early observations. JAMA 1966; 95:731–8.
(*21*) Habbema JD, Oortmarssen GJ, van Putten DJ, et al. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of a Health Insurance Plan of Greater New York Study. J Natl Cancer Inst 1986;77:317–20.
(*22*) Smart CR, Hartmann WH, Beahrs OH, Garfinkel L. Insights into breast cancer screening of younger women. Evidence from the 14-year follow-up of the Breast Cancer Detection Demonstration Project. Cancer 1993;72(4 Suppl):1449–56.
(*23*) Mettlin C, Smart CR. Breast cancer detection guidelines for women aged 40 to 49 years: rationale for the American Cancer Society reaffirmation of recommendations. CA Cancer J Clin 1994;44:248–55.
(*24*) Fletcher SW, Black W, Harris R, Rimer BK, Shapiro S. Report of the International Workshop on Screening for Breast Cancer. J Natl Cancer Inst 1993;85:1644–56.
(*25*) Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomized trials [published erratum appears in Lancet 1993;342:1372]. Lancet 1993;341:973–8.
(*26*) EUSOMA Evaluation Committee. Report of the European Society for Mastology Breast-Cancer Screening Evaluation Committee. J Eur Soc Mastology 1993;13:1–25.
(*27*) Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. JAMA 1995;273:149–54.
(*28*) Swanson GM, Ragheb NE, Lin CS, Hinkley BF, Miller B, Horn Ross PL. Breast cancer among black and white women in the 1980's: changing patterns in the United States by race, age, and extent of disease. Cancer 1993;72:788–98.
(*29*) Kopans DB, Moore RH, McCarthy KA, Hall DA, Hulka CA, Whitman GJ, et al. The positive predictive value of breast biopsy performed as a result of mammography: there is no abrupt change at age 50 years. Radiology 1996; 200:357–60.
(*30*) Lein BC, Alex WR, Zebley DM, Pezzi CM. Results of needle localized breast biopsy in women under age 50. Am J Surg 1996;171:356–9.
(*31*) Hermann G, Janus C, Schwartz IS, Papatestas A, Hermann DG, Rabinowitz JG. Occult malignant breast lesions in 114 patients: relationship to age and the presence of microcalcifications. Radiol 1988;169:321–4.
(*32*) Wazer DE, Gage I, Homer MJ, Krosnick SH, Schmid C. Age-related differences in patients with palpable breast carcinomas. Cancer 1996;78: 1432–7.
(*33*) Seidman H, Belb SK, Silverbert E. Survival experience with breast cancer detection demonstration project. CA 1987;37:258–90.
(*34*) Carter CL, Allen C, Hensson D. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer 1989;63:181–7.
(*35*) Wilhelm MC, Edge SB, Cole DD, DePardes E, Freierson HF. Nonpalpable invasive breast cancer. Ann Surg 1991;213:600–4.
(*36*) Haffty BG, Kornguth P, Fischer D, Beinfield M, McKhann C. Mammographically detected breast cancer: results with conservative surgery and radiation therapy. Cancer 1991;67:2801–4.

Note

The authors would like to thank Heidi Allen, Tumor Registrar at Roger Williams Medical Center, Elizabeth Angell for preparing the manuscript, and Martha Jamison for editorial assistance.

Increases in Ductal Carcinoma *In Situ* (DCIS) of the Breast in Relation to Mammography: A Dilemma

*Virginia L. Ernster, John Barclay**

The increased use of screening mammography has resulted in a marked increase in detected cases of ductal carcinoma *in situ* (DCIS) of the breast since the early 1980s. In 1993, there were an estimated 23,275 newly diagnosed cases of DCIS in the United States, of which 4,676 were in women aged 40–49. DCIS accounted for 14.7% of all newly diagnosed breast cancers in women aged 40–49 in 1993, and perhaps 40% of all mammographically detected breast cancers in this age group are DCIS. Among women aged 40–49, an estimated 1,890 mastectomies and 2,707 lumpectomies (with or without radiation) were performed for DCIS in 1993. There is an urgent need to better understand the relationship of mammographically detected DCIS to invasive and potentially life-threatening breast cancer. Better information about the appropriate treatment of DCIS is also needed to reduce the confusion and uncertainty many women and their physicians currently experience in the face of a DCIS diagnosis. For the present, women considering screening mammography should be told the likelihood of being diagnosed with DCIS and that only some DCIS cases may be clinically significant but almost all will be treated surgically. [Monogr Natl Cancer Inst 1997;22:151–156]

The widespread adoption of screening mammography has led to a marked increase in detected cases of ductal carcinoma *in situ* (DCIS) of the breast (*1*). DCIS is usually referred to as "preinvasive" or "noninvasive" cancer because it is confined to the milk ducts of the breast and has not spread to the surrounding breast tissue. Although DCIS lesions are usually not clinically palpable, they are visible on mammograms. Before the advent of mammography, they were often only detected incidental to a biopsy for a palpable lesion that was diagnosed as benign. Extrapolating data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program (*2*), we estimate that there were 23,275 newly diagnosed cases of DCIS in the United States in 1993, of which 4,676 were in women aged 40–49.

The increase in detected DCIS cases among women aged 40–49 is beneficial if those cases would have progressed to a life-threatening stage in the absence of screening at age 40–49. It is much less desirable if those cases rarely progress to invasive breast cancer (resulting in unnecessary treatment and anxiety) or if waiting until age 50 to be screened and to detect those cancers has no or only a minimal effect on the chance of breast cancer death. In short, we face the question of whether detecting DCIS through screening mammography, especially at ages 40–49, does more harm than good.

Based on follow-up of small numbers of untreated cases and of larger series of cases treated only by wide excision or lumpectomy, it appears that only a minority of DCIS cases will progress to or recur as invasive cancer (*29*). However, current knowledge of factors associated with recurrence is limited and, in the absence of good prognostic markers, treatment for DCIS is not radically different than that for Stage I invasive breast cancer. Thus, while screening mammography may benefit some women age 40–49 through early detection of potentially fatal breast cancers, it is potentially harming other women through detection of DCIS lesions that may be clinically insignificant but, for lack of better prognostic information, are almost always treated surgically. There is a critical need for better understanding of the epidemiologic, clinical, histopathologic, and genetic characteristics that distinguish those cases of DCIS that will go on to progress or recur from those that will not.

The fact that most abnormal mammography results are false positives has been discussed elsewhere (*3,4;* see also papers by Anderson, Lee, Sickles, Kerlikowske, Linver, Rimer, and Harris in this monograph). The focus here is on detection of what is technically a true positive (DCIS) but which, in some cases at least, may be clinically insignificant. For each woman who is contemplating screening, the willingness to risk a false positive or a positive result that may be clinically insignificant will differ, and it is therefore important that women know the probabilities of such outcomes in order to make their own informed decisions.

Trends in DCIS Incidence Rates

According to SEER data, between 1983 and 1993, age-adjusted incidence rates for DCIS in the United States increased 314%. (For comparison, the increase in incidence rates for invasive breast cancer over the same period was 15.7% percent.) Increases in DCIS incidence rates in the United States were dramatic for women 40 and older but much more modest for women under 40 (Fig. 1), who are much less likely to undergo screening mammography. In particular, among women 40–49

**Affiliation of authors:* Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco.

Correspondence to: Virginia L. Ernster, Ph.D., Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, CA.

See "Note" following "References."

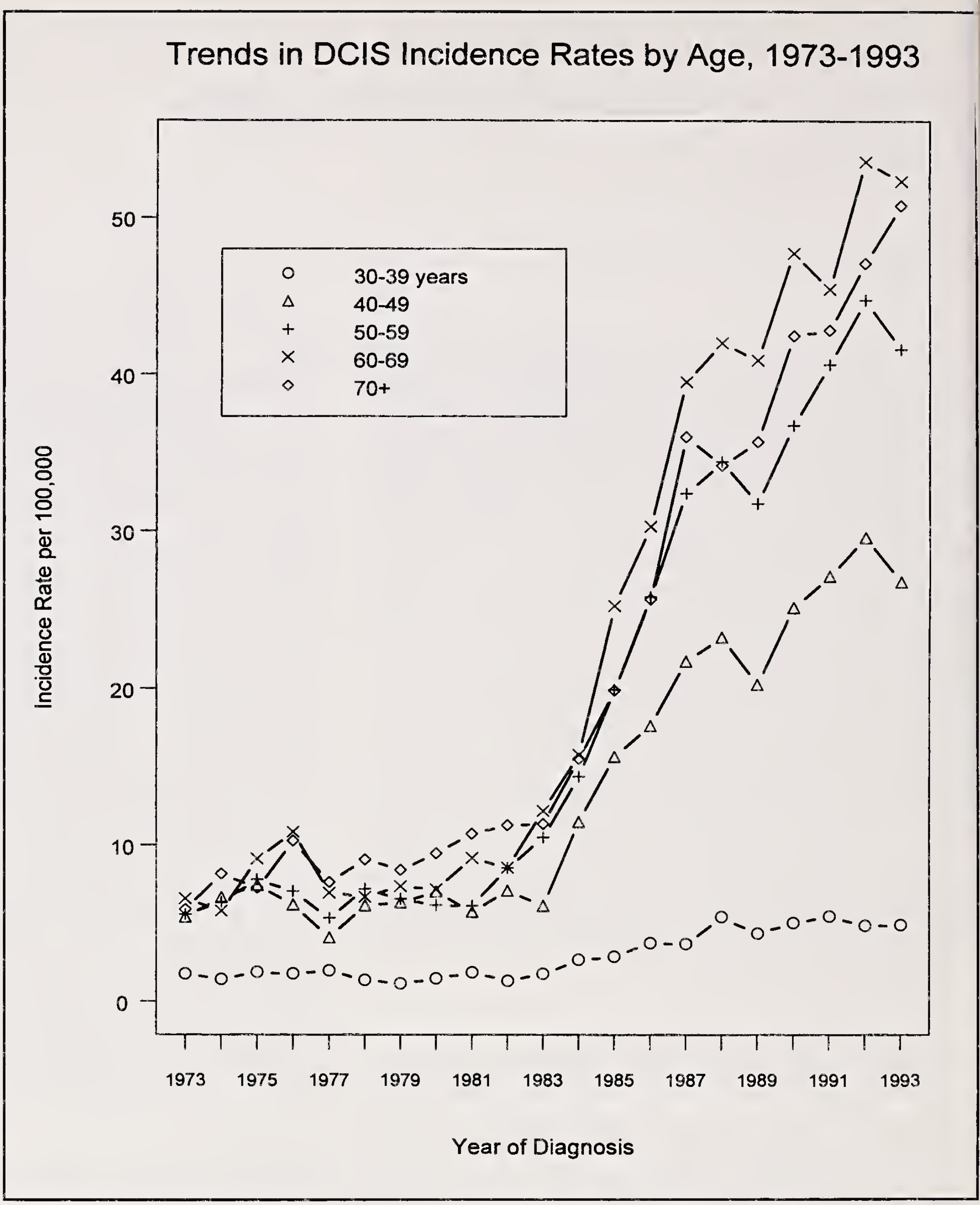

Fig. 1. Trends in DCIS incidence rates among U.S. women, by age, 1973–1993. Source: Calculated from data provided in (*2*).

years of age, incidence rates increased 339% between 1983 and 1993 (compared to 10.8% for invasive breast cancer). For all age groups combined, DCIS accounted for 2.8% of newly diagnosed breast cancers in the United States in 1973, 3.8% in 1983, and 12.5% in 1993. Among women ages 40–49, DCIS accounted for 3.7% of all breast cancers in 1973, 4.2% in 1983, and 14.7% in 1993 (*2*).

Relation of the DCIS Epidemic to Mammography Screening

There were 134 mammography machines in the United States in 1982 and an estimated 10,000 by 1990 (*5*). Meanwhile, use of mammography increased markedly; the proportion of U.S. women reporting recent mammography doubled between 1987 and 1992 (*6*). Most but not all of the increase in invasive breast cancer incidence during the 1980s has been attributed to increased detection through screening (*7*), and probably most of the excess of DCIS cases today compared to earlier years can be too.

Mammography screening programs, which focus on asymptomatic women, typically report much higher proportions of DCIS among all breast cancers detected than is observed in data from general tumor registries, which include cases in symptomatic women as well. For example, of breast cancers detected among women aged 40 and older undergoing first screening mammography at the Mobile Mammography Screening Program of the University of California, San Francisco (UCSF) during 1985–1996 who had no report of a palpable mass, 29.9% were DCIS. Among breast cancer cases detected among women aged 40–49 having their first mammograms in that program, 42.6% were DCIS, with substantial but lower proportions being

DCIS in the older age groups (Table 1). The higher proportion of DCIS among younger compared to older women with mammographically detected breast cancers is sometimes misinterpreted to mean that DCIS is more common in younger women and that screening is therefore particularly important for younger women. However, population-based cancer incidence data for the United States show that DCIS incidence rates do not decrease with age (except at the very oldest ages, when screening is less frequent), while rates for invasive breast cancer increase dramatically with age (*2*). Thus, DCIS comprises a higher proportion of mammographically detected cases in younger women not because they have more DCIS than older women, but because they have much less invasive breast cancer. Even in mammography screening programs, the number of DCIS cases detected per 10,000 first screening mammograms does not appear to be lower among women 50 and older than among women 40–49, as shown in Table 1 for the UCSF screening program. In sum, a larger proportion of the breast cancers detected among younger women are DCIS, which makes the issue of possible overdiagnosis of DCIS through screening especially important for this age group, but DCIS is not more common in women aged 40–49 than in older women.

Is DCIS a Precursor of Invasive Breast Cancer?

Is DCIS a precursor of invasive breast cancer and, if so, what proportion of DCIS progresses to invasive disease? To have actual proof that DCIS is a precursor of invasive breast cancer, we would have to diagnose but not treat a group of women with DCIS and follow them over time to determine whether invasive breast cancer occurs. Short of that, based on the more circumstantial evidence that we do have, it is probably safe to say that some fraction of DCIS progresses to clinically detectable invasive cancer, but DCIS is not an *obligate* precursor lesion.

Several lines of evidence support a precursor role. For example, the few epidemiologic studies that have examined risk factors for DCIS show similarities with invasive breast cancer, including a family history of breast cancer and nulliparity or older age at first childbirth (*8–11*). Although the role of menopausal hormone use in invasive breast cancer remains somewhat controversial, several (*11–13*) but not all (*14*) studies have shown it to be a risk factor for DCIS. Secondly, many laboratory studies have compared genetic markers in DCIS and invasive breast tumors and found similarities, much more commonly with high-grade or comedo DCIS than low-grade or non-comedo DCIS (*15–17*), although whether these markers actually predict DCIS progression to invasive disease is unclear. Finally, the distribution of DCIS lesions in the breast is almost identical to the distribution of invasive breast cancers (*18*).

Table 1. Percent of breast cancer that is DCIS among cancers detected during first screening mammography, and cases of DCIS and invasive cancer detected per 10,000 first screenings*, by age, UCSF Mobile Mammography Screening Program, 1985–1996**

			Cases per 10,000 screens	
Age	Total cancers	% DCIS	DCIS	Invasive cancer
30–39	11	90.9%	11	1
40–49	47	42.6%	14	19
50–59	56	28.6%	20	51
60–69	58	19.0%	22	94
70+	39	28.2%	42	106

*Women who present for screening with no report of a palpable mass.

**Based on the authors' analysis of the database of E. A. Sickles, M.D., Department of Radiology, University of California, San Francisco.

On the other hand, there is evidence to suggest that perhaps the majority of DCIS cases do not progress to clinically significant invasive breast cancer and therefore might not be considered precursor lesions. For example, the prevalence of DCIS in seven autopsy series of women who died of causes unrelated to breast cancer ranges from as low as 0.2% to as high as 18.2%, with four of the studies finding a prevalence of greater than 10% (*19–25*). There are also several series of women who had breast biopsies 30 to 40 years ago that were interpreted at the time to be benign breast disease and who were therefore untreated beyond biopsy. When their pathology slides were re-reviewed some years later, it was determined that the women actually had had DCIS, and they were then followed to determine what proportion subsequently developed invasive breast cancer. Although these studies are often cited in support of a precursor role for DCIS, even they suggest that following biopsy alone, the majority of DCIS does not progress to invasive breast cancer. Perhaps the two most informative series are those of Page *et al.* and of Eusebi *et al.*, as other studies generally had smaller numbers or large losses to follow-up. Page *et al.* identified 28 women who were biopsied between 1952 and 1968 in Nashville, Tennessee, and who had an average of 30 years of follow-up, during which time nine women (32%) developed ipsilateral invasive breast cancer (*26,27*). Eusebi *et al.*, identified 80 patients who were biopsied between 1964 and 1976 in northern Italy, with an average of 17.5 years of follow-up; nine of these women (11.3%) developed ipsilateral invasive breast cancer (*28*). Even these are small clinical series; also, it is difficult to know the extent to which we can extrapolate from the experience of these historic cases of DCIS, which occurred in women with breast symptoms, to the experience of women diagnosed with DCIS today, most of which is occult disease detected mammographically.

There are also a number of series of women treated by wide excision alone who have been followed for breast cancer recurrence (either as DCIS or invasive disease). Most of these studies show that about 4% to 5% of such cases recur as invasive cancer after 3 to 10 years of follow-up (*29*). The best known randomized trial of DCIS treatment (lumpectomy versus lumpectomy plus radiation) reported a 10.5% five-year cumulative incidence of ipsilateral invasive cancer in women treated by lumpectomy alone (*30*). The lower recurrence rates in the earlier case series may reflect greater selection for smaller tumors in those studies compared to randomized trials.

It is generally thought that patients with specific histologic types of DCIS, namely those with high nuclear grade or comedo-type DCIS, are at greatest risk of recurrence, although whether this holds up after long-term follow-up or whether nuclear grade or comedo-type DCIS affects actual survival per se is unclear (*31*). The Eastern Cooperative Oncology Group (ECOG) has proposed a large observational study of minimal treatment (local excision) for DCIS of small size and low nuclear grade, which would provide useful information on recurrence for women with what is currently considered to be low-risk DCIS.

We do know that the vast majority of women with DCIS do quite well in terms of subsequent breast cancer mortality. Among women in the population-based SEER cancer database who were diagnosed with DCIS between 1978 and 1993, 0.5% died of breast cancer within five years and 2.6% within 10 years (*32*). Whether these low proportions reflect the effectiveness of treatment (almost all cases were treated surgically) or the fact that DCIS is a relatively benign disease to begin with—or both—is unclear. One caveat is that the experience of women classified in the SEER database as having DCIS may overestimate the likelihood of breast cancer death associated with DCIS, since reexamination of original pathology reports for those women suggests that up to 15% of cases coded as DCIS in SEER actually had early invasive cancer (Ann Coleman, Ph.D., personal communication). Moreover, the women with the longest follow-up are those diagnosed in the era preceding the widespread adoption of screening mammography, and it may be inappropriate to extrapolate from their experience to that of women diagnosed more recently by mammography. Although we still know relatively little about the natural history of DCIS, it is probably fair to conclude that some DCIS cases will progress to clinically significant invasive breast cancer but many will not.

DCIS Treatment Trends: Mastectomy versus Lumpectomy

Most all DCIS is treated surgically, either by mastectomy or by lumpectomy with or without radiation; according to SEER data for 1993, only 1.7% of DCIS cases did not have surgery. As shown in Table 2 for women of all ages combined, the proportion of DCIS cases treated by mastectomy has declined substantially over time, from 71% in 1983 to 39.7% in 1993; among women aged 40–49 years, the decline over that time period was from 75.8% to 40.4%. The proportion of DCIS cases treated by breast conserving therapy has increased over time; among women aged 40–49 in 1993, 32.9% of eases were treated by lumpectomy plus radiation and 25% by lumpectomy alone (*2*). Extrapolating from SEER incidence rates and treatment patterns to the general U.S. population, an estimated 9,245 mastectomies were performed for DCIS in the United States in 1993, of which 1,890 were in women 40–49; an additional 2,707 women aged 40–49 are estimated to have had lumpectomy with or without radiation in 1993. Over the period 1983–1993, an estimated 89,845 breasts were removed for DCIS in U.S. women, including 17,456 among women aged 40–49, and presumably most of those cases were mammographically detected.

Although over half of DCIS cases in the United States were treated by mastectomy until 1991, it is of interest that there are no randomized clinical trials of mastectomy versus other treatment options for DCIS, nor are there ever likely to be, given that lumpectomy plus radiation already has been shown to be equal in effectiveness to mastectomy for treatment of early-stage invasive breast cancer. The largest randomized clinical trial of DCIS treatment published to date is the National Surgical Adjuvant Breast Project study (NSABP-B-17); it randomized approximately 800 women with DCIS to receive either lumpectomy alone or lumpectomy plus radiation. Results published in 1993, based on a mean follow-up of 43 months, showed statistically significantly lower rates of breast cancer recurrence in the group that received lumpectomy plus radiation (*30*). Recently presented data based on eight years of follow-up continue to confirm the difference: invasive breast cancer had occurred in 13.4% of the women treated by wide excision alone compared to 3.9% of those treated by lumpectomy plus radiation (Norman Wolmark, M.D., "The NSABP Experience in DCIS," 19th Annual San Antonio Breast Cancer Symposium, San Antonio, Texas, Dec. 11, 1996). However, although numbers of deaths were small (only about 20 deaths from all causes combined in each group), survival per se did not differ between the two groups. Other randomized clinical trials of DCIS treatment by lumpectomy with or without radiation or tamoxifen are underway (*33*).

Current Dilemmas Posed by Detection of DCIS

Our increased ability to detect DCIS through mammography and the resultant "epidemic" of reported DCIS cases present women and their physicians with a dilemma: probably only a minority of DCIS cases will actually go on to invasive breast cancer and become clinically important. However, since current medical knowledge does not permit us to identify which women

Table 2. Estimated numbers of DCIS cases, percent treated by mastectomy, and estimated numbers of mastectomies for DCIS in all U.S. women and in women aged 40–49, 1983–1993

	Estimated number of DCIS cases		% Cases treated by mastectomy		Estimated number of mastectomies	
Year	All women	Ages 40–49	All women	Ages 40–49	All women	Ages 40–49
1983	4,901	742	71.0	75.8	3,479	563
1984	7,069	1,433	66.6	67.7	4,706	971
1985	9,897	1,991	59.5	57.5	5,887	1,144
1986	12,279	2,283	56.1	57.0	6,890	1,300
1987	16,034	3,000	59.3	62.8	9,515	1,884
1988	17,196	3,345	57.8	56.3	9,934	1,882
1989	16,584	3,086	56.3	53.0	9,334	1,635
1990	19,890	3,970	53.7	51.0	10,682	2,025
1991	20,735	4,325	47.8	44.2	9,908	1,912
1992	23,438	4,973	43.8	45.3	10,265	2,250
1993	23,275	4,676	39.7	40.4	9,245	1,890
Total	171,298	33,824			89,845	17,456

*Based on extrapolations from NCI's SEER program data on cancer incidence rates and treatment patterns (*2*).

with DCIS will progress to invasive cancer and which will not, at present most women with a DCIS diagnosis are treated surgically. The hope is that by detecting malignant changes as early as possible, we are saving lives. The concern is that we may be detecting changes which for many women would never become life threatening or even clinically apparent and that, in the process, we are overtreating women. Thus, it behooves us to learn whether there is a survival benefit associated with early detection and treatment of DCIS and, if so, whether it obtains only for specific subtypes of DCIS. We need to know the appropriate clinical strategies for different subtypes of DCIS. These strategies could range from biopsy followed by watchful waiting, on the one hand, to mastectomy on the other. The situation is similar to the current dilemma posed by prostate specific antigen (PSA) screening for prostate cancer; while debate continues as to whether that test reduces risk of prostate cancer death, it is known that PSA screening picks up many occult cancers that are clinically unimportant but for which thousands of men have had their prostates removed, in some cases resulting in impotence and incontinence (*34*).

On the basis of breast cancer incidence rates and actual treatment patterns in the SEER data (*2*), we have estimated the numbers of surgeries for breast cancer among U.S. women aged 40–49 in 1983 and 1993 by stage of disease, assuming that the total number of women in the population was the same in both years (i.e., we used the 1993 population data). The total number of breast surgeries for breast cancer among women 40–49 increased from 24,343 to 30,535 between 1983 and 1993 (Table 3). Among women 40–49, there were increases in breast surgeries of 333% for DCIS and 32% for localized invasive cancer, and decreases of 8% for regional disease and 4% for distant disease. Because the proportion of cases treated by mastectomy declined over time for all stages of breast cancer, there were fewer mastectomies performed overall in 1993 than in 1983; however, as we have seen earlier, this was not true for DCIS because the dramatic increase in DCIS incidence rates resulted in an increase in the number of DCIS-related mastectomies, despite the declining proportion of DCIS cases being treated by mastectomy over time. Thus, we have a fairly good idea of the likelihood of a DCIS diagnosis for women undergoing mammography screening and of the likelihood of various types of breast surgery. What we still don't know is whether detection of breast cancer at the DCIS stage ultimately saves lives.

Table 3. Estimated numbers* of breast surgeries for DCIS and other stages of breast cancer among white and black U.S. women aged 40–49, 1983 and 1993**

Stage and type of surgery	1983 Estimated number	1983 Percent of all cases	1993 Estimated number	1993 Percent of all cases
DCIS				
Breast conserving	258	24	2,704	58
Mastectomy	804	76	1,890	40
Total	1,062	100	4,594	98
Localized				
Breast conserving	2,442	20	8,622	55
Mastectomy	9,373	78	6,996	44
Total	11,815	98	15,618	99
Regional				
Breast conserving	1,289	13	2,980	32
Mastectomy	8,690	85	6,238	67
Total	9,979	98	9,218	99
Distant				
Breast conserving	167	13	222	18
Mastectomy	698	56	609	40
Total	865	69	831	67
All stages combined				
Breast conserving	4,338	17	14,619	46
Mastectomy	20,005	78	15,916	50
Total	24,343	95	30,535	96

*Based on extrapolations from NCI's SEER program data on cancer incidence rates and treatment patterns (*2*).

**Assumes the same number of women in the population in both years, namely the population distribution of U.S. women in 1993. Estimates based on rates for *in situ* cancer not including cases of lobular carcinoma *in situ*. Estimated numbers within each stage category do not include cases with no surgery or those with breast cancer-related surgery outside of the breast. However, the percentages shown reflect the proportions of all cases in each stage category (including those with no surgery or those with breast cancer-related surgery outside of the breast) that were treated by breast-conserving surgery or mastectomy, and those proportions usually do not add up to 100% of all cases in the category. The "All stages combined" category includes breast cancers of unknown stage. If cases with no surgery and those with breast cancer-related surgery outside of the breast had been included, numbers would be slightly higher (e.g., totals would be 25,511 and 31,618 for 1983 and 1993, respectively).

Directions for Future Research

It is agreed that most of the increase in reported cases of DCIS results from better detection of the disease through mammography rather than a true excess of new cases. Especially given the numbers of women diagnosed with DCIS in recent years, there is urgent need to better understand the relationship of mammographically detected DCIS to invasive and potentially life-threatening breast cancer. DCIS shares at least some risk factors and genetic changes in common with invasive breast cancer, which suggests etiologic similarities and supports the position that at least some DCIS cases are precursors to invasive disease. Other evidence suggests that many cases of DCIS are not clinically significant; in most autopsy series examined, occult DCIS is not uncommon in women who died of causes other than breast cancer, and small historical series of women with DCIS who received no treatment beyond diagnostic biopsy show that most did not subsequently develop clinically apparent invasive breast cancer. Thus, biologic and epidemiologic studies are needed to identify prognostic markers and risk factors associated with progression; these studies should focus on specific histologic types of DCIS and perhaps correlate them with breast imaging studies.

Better information about the appropriate treatment of DCIS is also needed to reduce the confusion and uncertainty many women and their physicians currently experience in the face of a DCIS diagnosis. For the present, informed decision making about screening mammography should include the likelihood of being diagnosed with DCIS, with an explanation that only some DCIS cases may be clinically significant, as well as the likelihood of having breast surgery as a result of DCIS detection.

References

(*1*) Ernster VL, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996; 275:913–8.

(*2*) Surveillance, Epidemiology, and End Results (SEER) Program public use CD-ROM (1973–1993). Bethesda (MD): National Cancer Institute, DCPC, Surveillance Program, Cancer Statistics Branch, 1996.

(3) Harris R, Leininger L. Clinical strategies for breast cancer screening: weighing and using the evidence. Ann Intern Med 1995;122:539–47.
(4) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.
(5) Brown ML, Kessler LG, Rueter FG. Is the supply of mammography machines outstripping need and demand? An economic analysis. Ann Intern Med 1990;113:547–52.
(6) Trends in cancer screening-United States, 1987 and 1992. MMWR Morb Mortal Wkly Rep 1996;45:57–61.
(7) Wun LM, Feuer EJ, Miller BA. Are increases in mammographic screening still a valid explanation for trends in breast cancer incidence in the United States? Cancer Causes Control 1995;6:135–44.
(8) Newcomb PA, Storer BE, Marcus PM. Risk factors for carcinoma in situ of the breast [abstract]. Cancer Epidemiol Biomark Prev 1994;3:189.
(9) Weiss HA, Brinton LA, Brogan D, Coates RJ, Gammon MD, Malone KE, et al. Epidemiology of in situ and invasive breast cancer in women aged under 45. Br J Cancer 1996;73:1298–305.
(10) Kerlikowske K, Barclay J, Grady D, Sickles EA, Ernster VL. Comparison of risk factors for ductal carcinoma in situ and invasive breast cancer. J Natl Cancer Inst 1997;89:76–82.
(11) Longnecker MP, Bernstein L, Paganini-Hill A, Enger SM, Ross RK. Risk factors for in situ breast cancer. Cancer Epidemiol Biomarkers Prev 1996; 5:961–5.
(12) Brinton LA, Hoover R, Fraumeni JF Jr. Menopausal oestrogens and breast cancer risk: an expanded case-control study. Br J Cancer 1986;54:825–32.
(13) Schairer CA, Byrne C, Keyl PM, Brinton LA, Sturgeon SR, Hoover RN. Menopausal estrogen and estrogen-progestin replacement therapy and risk of breast cancer (United States). Cancer Causes Control 1994;5:491–500.
(14) Stanford JL, Weiss NS, Voight LF, Daling JR, Habel LA, Rossing MA. Combined estrogen and progestin hormone replacement therapy in relation to risk of breast cancer in middle-aged women. JAMA 1995;274:137–42.
(15) Pallis L, Skoog L, Falkmer U, Wilking N, Rutqvist LE, Auer G, et al. The DNA profile of breast cancer in situ. Eur J Surg Oncol 1992;18:108–11.
(16) Radford DM, Fair KL, Phillips NJ, Ritter JH, Steinbrueck T, Holt MS, et al. Allelotyping of ductal carcinoma in situ of the breast: deletion of loci on 8p, 13q, 16q, 17p and 17q. Cancer Res 1995;55:3399–4405.
(17) O'Connell P, Pekkel V, Fuqua S, Osborne CK, Allred DC. Molecular genetic studies of early breast cancer evolution. Breast Cancer Res Treat 1994;32:5–12.
(18) Surveillance, Epidemiology, and End Results (SEER) Program public use CD-ROM (1973–1992). Bethesda (MD): National Cancer Institute, DCPC, Surveillance Program, Cancer Statistics Branch; 1995.
(19) Wellings SR, Jensen HM. On the origin and progression of ductal carcinoma in the human breast. J Natl Cancer Inst 1973;50:1111–8.
(20) Kramer WM, Rush BF Jr. Mammary duct proliferation in the elderly. A histopathologic study. Cancer 1973;31:130–7.
(21) Nielsen M, Jensen J, Andersen J. Precancerous and cancerous breast lesions during lifetime and at autopsy. A study of 83 women. Cancer 1984; 54:612–5.
(22) Alpers CE, Wellings SR. The prevalence of carcinoma in situ in normal and cancer-associated breasts. Hum Pathol 1985;16:796–807.
(23) Bhathal PS, Brown RW, Lesueur GC, Russell IS. Frequency of benign and malignant breast lesions in 207 consecutive autopsies in Australian women. Br J Cancer 1985;51:271–8.
(24) Nielsen M, Thomsen JL, Primdahl S, Dyreborg U, Andersen JA. Breast cancer and atypia among young and middle-aged women: a study of 110 medicolegal autopsies. Br J Cancer 1987;56:814–9.
(25) Bartow SA, Pathak DR, Black WC, Key CR, Teaf SR. Prevalence of benign, atypical, and malignant breast lesions in populations at different risk for breast cancer. A forensic study. Cancer 1987;60:2751–60.
(26) Page DL, Dupont WD, Rogers LW, Landenberger M. Intraductal carcinoma of the breast: follow-up after biopsy only. Cancer 1982;49:751–8.
(27) Page DL, Dupont WD, Rogers LW, Jensen RA, Schuyler PA. Continued local recurrence of carcinoma 15–25 years after a diagnosis of low grade ductal carcinoma in situ of the breast treated only by biopsy. Cancer 1995; 76:1197–1200.
(28) Eusebi V, Foschini MP, Cook MG, Berrino F, Azzopardi JG. Long-term follow-up of in situ carcinoma of the breast with special emphasis on clinging carcinoma. Semin Diagn Pathol 1989;6:165–73.
(29) Hetelekidis S, Schnitt SJ, Morrow M, Harris JR. Management of ductal carcinoma in situ. CA Cancer J Clin 1995;45:244–53.
(30) Fisher B, Costantino J, Redmond C, Fisher E, Margolese R, Dimitrov N, et al. Lumpectomy compared with lumpectomy and radiation therapy for the treatment of intraductal breast cancer. N Engl J Med 1993;328:1581–6.
(31) Fisher ER. Pathobiological considerations relating to the treatment of intraductal carcinoma (ductal carcinoma in situ) of the breast. CA Cancer J Clin 1997;47:52–64.
(32) Ernster VL, Barclay J, Ballard-Barbash R, Kerlikowske K, Wilkie H. Mortality from breast cancer and relative survival among women with DCIS: an examination of population-based SEER data for 1978–1991 [abstract]. Breast Cancer Res Treat 1996;41:238.
(33) van Dongen JA, Holland R, Peterse JL, Fentiman IS, Lagios MD, Millis RR, et al. Ductal carcinoma in-situ of the breast; second EORTC consensus meeting. Eur J Cancer 1992;28:626–9.
(34) Woolf SH. Screening for prostate cancer with prostate-specific antigen. An examination of the evidence. N Engl J Med 1995;333:1401–5.

Note

This work was undertaken while the authors were supported in part by NCI grants CA 58207–05 (SPORE) and U01 CA 63740–02.